

PCT/JP 03/07514

日本国特許庁

JAPAN PATENT OFFICE

30.07.03 #2

10/51754A

REC'D 19 SEP 2003

WIPO PCT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日  
Date of Application: 2002年 6月12日

出願番号  
Application Number: 特願2002-171851  
[ST. 10/C]: [JP 2002-171851]

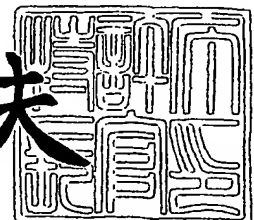
出願人  
Applicant(s): 理化学研究所  
株式会社ダナフォーム

**PRIORITY DOCUMENT**  
SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH  
RULE 17.1(a) OR (b)

2003年 9月 4日

特許庁長官  
Commissioner,  
Japan Patent Office

今井康夫



BEST AVAILABLE COPY

出証番号 出証特2003-3072107

【書類名】 特許願  
【整理番号】 02771  
【特記事項】 特許法第 3 6 条の 2 第 1 項の規定による特許出願  
【あて先】 特許庁長官殿  
【国際特許分類】 C12N 15/00  
【発明者】

【住所又は居所】 茨城県つくば市稲荷前 2 2 - 1 - 2 0 1

【氏名】 林崎 良英

【発明者】

【住所又は居所】 埼玉県和光市広沢 2 - 1 理化学研究所内

【氏名】 カルニンチ ピエロ

【発明者】

【住所又は居所】 東京都港区三田一丁目 3 - 3 5 株式会社ダナフォーム  
内

【氏名】 マティアス・ティ・ハーベス

【特許出願人】

【識別番号】 000006792

【氏名又は名称】 理化学研究所

【特許出願人】

【識別番号】 502049114

【氏名又は名称】 株式会社ダナフォーム

【代理人】

【識別番号】 100088546

【弁理士】

【氏名又は名称】 谷川 英次郎

【電話番号】 03(3238)9182

【手数料の表示】

【予納台帳番号】 053235

【納付金額】 35,000円

【提出物件の目録】

【物件名】 外国語明細書 1

【物件名】 外国語図面 1

【物件名】 外国語要約書 1

【プルーフの要否】 要

**【書類名】** 外国語明細書

1 Title of Invention Method to utilize 5' ends of transcribed regions for cloning and analysis

2 Claims

1. A method to prepare nucleic acids tags corresponding to the 5' end of transcribed regions said mRNA.

2. A method according to claim 1 where concatamers of such 5' end tags are produced.

**【請求項 2】** A method in which such 5' end specific sequence tags derived from transcribed regions said mRNA are analyzed by sequencing.

**【請求項 3】** The method for preparing concatemers of a plurality of at least two or more nucleic acid fragments having information on nucleotide sequences of 5' end regions of a plurality of nucleic acids related to transcribed in a sample, comprising

A first step of selectively collecting a plurality of cDNAs containing regions complementary to 5'-end regions of mRNAs, which cDNAs are formed by using RNA or mRNAs derived from a biological sample or in vitro synthesized RNA derived from cDNA - or tag - libraries in the sample as templates;

A second step to collecting fragments containing cDNA regions including at least the regions corresponding to the 5'-end regions of said mRNAs or cDNA;

And a third step of creating a concatamer of such 5' end nucleic tags.

**【請求項 4】** The method as in claim 4 but in which  
The first step is substituting the cap-structure of mRNAs with an oligonucleotide;

The second step constitute in the formation of full-length cDNA;

The third step involve cleavage of a 5' end tag and formation of concatamers.



【請求項 5】 The method according to claim 4, wherein said first step comprises the steps of synthesizing the first-strand cDNAs using mRNAs as templates; attaching a selective binding substance to the cap structures of said mRNAs; cleaving single-stranded RNAs; binding said selective binding substance to a corresponding selective binding substance immobilized on a support, which corresponding selective binding substance selectively binds to said selective binding substance; and recovering said cDNA.

【請求項 6】 A method as in claim 4 where the first step to isolate the full-length cDNA includes an RNase digestion step followed by treatment with an immobilized cap-binding substance followed by eluting such full-length cDNAs.

【請求項 7】 A method to add a sequence connected to the 5' end a nucleic acids corresponding to the 5' terminal part of a transcript, when such that can be recognized by a substance that is capable of cleaving such nucleic acids outside the recognition sequence.

【請求項 8】 The method according to claim 4, wherein said selective binding substance is biotin, and said corresponding selective binding substance is avidin, streptavidin or an avidin or streptavidin derivative which specifically binds to biotin.

【請求項 9】 The method according to the claim 4 where the selective binding substance is digoxigenin and said corresponding binding substance is an antibody directed against digoxigenin.

【請求項 10】 A method according to claim 4 or 9, wherein a selective binding substance is bound to a corresponding selective binding substance which is immobilized on to a support, and where such a support is made of magnetic beads, agarose beads, or latex beads

【請求項 11】 The method according to any one of claims 4 and 6 to 11, wherein said second step comprises the steps of binding a linker ha

ving at least a restriction site for a substance that cleaves DNA outside its recognition sequence in the end region corresponding to the 5' end of said nucleic acids corresponding to the 5' end of genes, and a random oligomer region at the 3' end region; synthesizing a second-strand cDNA using said linker or other oligonucleotides partially or totally corresponding to the linker as a primer and said cDNA as a template; treating the obtained linker-bound double-stranded cDNA with said restriction enzyme; and selectively recovering fragments yielded by cleavage by the restriction enzyme, which fragments contain said linker moieties and part of 5' end cDNA.

【請求項 12】 The method according to any of claims 4 to 12, wherein a selective binding substance is attached to said linker; and the step of selectively recovering said fragments containing said linker moieties comprises the steps of binding said selective binding substance to a corresponding selective binding substance immobilized on a support, which corresponding selective binding substance selectively binds to said selective binding substance; and recovering said support.

【請求項 13】 The method according to any of claims 4 to 13, wherein said selective binding substance is biotin, and said corresponding selective binding substance is avidin, streptavidin, or an avidin derivative or derivatives of streptavidin which specifically binds to biotin.

【請求項 14】 The method according to any of claims 4 to 13 where the selective binding substance is digoxigenin and said corresponding binding substance is an antibody directed against digoxigenin.

【請求項 15】 The method according to any one of claims 4 to 15, wherein said restriction enzyme is a substance with an enzymatic activity to recognize nucleic acid and to cleave at a site different from the recognition site.

【請求項 16】 The method according to any one of claims 4 to 16, w

herein said restriction enzyme is a class II restriction enzyme like Gsu I, MmeI, Bpm I or Bsg I

【請求項 17】 The method using nucleic acid fragments obtained according to any one of claims 1 to 18, for further comprising the steps of cloning into concatamer.

【請求項 18】 A method for determining nucleotide sequences of 5'-end regions of a plurality of mRNAs by sequencing said concatemer prepared by the method according to any one of claims 1 to 18.

【請求項 19】 A method, which is the same method according to any one of claim 1 to 18, except that preliminarily obtained cDNAs having complete length is used instead of carrying out said first step.

【請求項 20】 A method to produce 5' end nucleic acids tags corresponding to the 5' ends of mRNA, in which a mixture of RNA molecules is prepared from a preexisting full-length cDNA library and the obtained RNA carries at the 5' end of the RNAs a sequence cleavable by a substance able to recognize a nucleic acids and cleave outside its recognition sequence.

【請求項 21】 A method to produce 5' end nucleic acids tags corresponding to the 5' ends of mRNA, in which a mixture of nucleic acids TAG molecules is prepared from a preexisting full-length cDNA library carrying close to the 5' end of a sequence cleavable by a substance able to recognize a nucleic acids and cleave outside its recognition sequence, which is used to produce a nucleic acid TAG molecule.

【請求項 22】 The concatemer prepared by the method according to any one of claims 1 to 22.

【請求項 23】 A vector comprising said concatemer according to claim 23.

【請求項 24】 A sequence, which is derived from a concatemer prepared by the method according to any one of claims 1 to 22.

【請求項 25】 A method based on any of claims 1 to 22, which allows to determine the transcriptional status of a given cell and therefore the transcriptional networking.

【請求項 26】 A method, which is the same method according to any one of claims 1 to 22 to obtain expression data on a plurality of mRNA or cDNA in a sample.

【請求項 27】 A method, which is the same method according to any one of claims 1 to 22, to quantify expression data on a plurality of mRNA A in a sample.

【請求項 28】 A method, which uses sequence information obtained from concatemers prepared by a method according to any one of claims 1 to 22, is used to build a database holding sequence information derived from the concatemers.

【請求項 29】 A method, which is the same method according to any one of claims 1 to 22, to identify open reading frames in a genomic sequence said genome.

【請求項 30】 A method, which is the same method according to any one of claims 1 to 22, to identify start sites of transcription and regulatory sequences upstream of the start site of transcription in a genomic sequence said genome.

【請求項 31】 A method, which uses sequence information obtained from concatemers prepared by a method according to any one of claims 1 to 22, to clone a full-length or partial cDNA from a plurality of nucleic acids.

【請求項 32】 A method, which uses sequence information obtained from concatemers prepared by a method according to any one of claims 1 to 22, to analyze the activity of regulatory regions in a genome said promoter.

【請求項 33】 A method, which uses sequence information obtained from

rom concatemers prepared by a method according to any one of claims 1 to 22, to inactivate a gene.

【請求項 34】 A method, which uses sequence information obtained from concatemers prepared by a method according to any one of the claims 1 to 22, to synthesis nucleotide sequences said linker.

【請求項 35】 A method, which uses sequence information obtained from concatemers prepared by a method according to any one of the claims 1 to 22, to synthesize nucleotide sequences said primers.

【請求項 36】 A method, which uses sequence information obtained from concatemers prepared by a method according to any one of the claims 1 to 22, to obtain extended nucleotide sequences derived from the 5'-ends of transcripts said sequencing.

【請求項 37】 A method according to any one of claims 1 to 8, wherein a single stranded cDNA is ligated to a double stranded synthetic oligonucleotide said linker, wherein the linker has a single stranded overhang encompassing a nucleotide sequence said tag, which was obtained from concatemers prepared by a method according to any one of the claims 1 to 13, wherein the linker is attached to a selective binding substance and the selective binding substance is attached to a corresponding selective binding substance said support, and where such linker bound to the support is used to enrich a specific nucleotide sequence said 1<sup>st</sup> strand cDNA said RNA transcript.

【請求項 38】 A method according to any one of claims 1 to 8, wherein a single stranded cDNA is ligated to a double stranded linker said primer, where a selective binding substance is attached to said linker, and where selectively binding substance is attached to a corresponding selective binding substance said support, and where such DNA template is used to obtain the nucleotide sequences of the 5'-region of an initial transcript said RNA.

【請求項 39】 A method based on any of the claims 1-39 to be used for the development of diagnostic tools.

### 3 Detailed Description of Invention

#### 【0001】

#### 【発明の属する技術分野】

The present invention relates to a method to selectively collect multiple nucleic acid fragments containing information on the nucleotide sequences at the 5' end site of multiple mRNAs within a sample. The method of the present invention is effective for analyzing the mRNAs contained within the sample, for discovering new genes, and for studies on gene regulation.

#### 【0002】

#### 【従来技術】

To utilize genomic information parts of the genome are transcribed into mRNA. For the understanding of the genome and its use in regulatory processes, information on individual mRNA species is required, which should include their partial or full-length nucleotide sequence and their relative or absolute quantity in a given biological context.

#### 【0003】

Conventionally, the base sequences in mRNAs contained in a cell or tissue sample had been analyzed by preparing a cDNA library by reverse transcription, using mRNAs as templates and investigating the individual insert cDNA fragments within said cDNA library. Since a sample contains a large number of varied mRNAs, the conventional method is of limited efficiency to analyze gene expression profiles and to identify rare genes. Therefore other technologies have been invented to monitor the expression patterns of mRNA in complex samples and to identify genes by short sequence elements said tags.

#### 【0004】

High-throughput expression profiling is commonly performed by the use of so-called DNA microarrays (Jordan B., DNA Microarrays: Gene Expression Applications, Springer-Verlag, Berlin Heidelberg New York, 2001; Schena A, DNA Microarrays, A Practical Approach, Oxford University Press, Oxford 1999). For such experiments specific probes representing individual genes or transcripts are placed on a support and simultaneously hybridized with a plurality of samples. Positive signals will be obtained where a probe on the support reacts with a molecule presented with the sample. These experiments allow the parallel analysis of a large number of genes or transcripts. However, the approach is limited to the fact that only genes or transcripts can be studied, which were initially identified by other experimental means. Such means can include cDNA libraries, partial sequence tags and/or results obtained from computer predictions. Due to the limitations of DNA microarray experiments alternative approaches are in use for gene discovery and expression profiling, which are based on partial sequences said tags obtained from a plurality of mRNA samples.

#### 【 0 0 0 5 】

The so-called SAGE (Serial Analysis of Gene Expression) method is known as an efficient method of obtaining partial information on the base sequences in mRNAs (Velculescu V.E. et al., Science 270, 484-487 (1995)). This method forms DNA concatamers by ligating multiple short DNA fragments (about 10 bp) containing information on the base sequences at the 3' end site of multiple mRNAs, and determines the base sequences in these DNA concatamers. It is a method for finding out partial information on the base sequences at the 3' end site of multiple mRNAs. When only a short base sequence close to the 3' end is available but the mRNAs itself is already known, the SAGE method can often identify the mRNA, although the available base sequence is as short as about 10 bp. This method is currently in wide use as an important method for analyzing genes expressed

in specific cells or tissues.

【 0 0 0 6 】

【発明が解決しようとする課題】

While the SAGE method can be used to learn a partial base sequence at the 3' end site of mRNAs, it is difficult to clone new genes based on the information in such short sequences at the 3' end site alone. Despite the application, SAGE does not teach how to obtain cDNA clones close to the 5' end of the cDNA. In fact, 4 bp restriction enzymes of class IIs are used. A 4bp cutter usually cleaves on average a few hundred nucleotides, which is on average  $1/10^{\text{th}}$  of the average size of an mRNA transcript. Thus SAGE principles strongly suggest that 3' ends are collected with high prevalence, and no information can be collected about the 5' end for most of the transcript. In addition 10 bp tags have often been insufficient for specific gene identification and mapping to genomic sequences said entire or partial genomes. Therefore, the 10 bp tags are used to identify only a "sage-tag", which comprises a part of a mRNA. Notice that mammalian mRNA comprises only 3-5% of the transcribed part of the mammalian genome and the specific "sage-tag" comprises a subfraction of this 3-5%, which lies in proximity of the class IIs restriction enzyme used in the analysis. Since a 4 bp restriction enzyme cuts approximately a random sequence every  $4^4$  bp (256 bp), the "sage-tags" can represent approximately  $1/256$  of the 3-5% expressed fraction of the genome (calculation=less than 0.02%). Therefore, the SAGE techniques teach that essentially it is not possible to use SAGE-tags to analyze a genome but only a very limited fraction of it.

【 0 0 0 7 】

Accordingly, the invention claimed in this application aims to provide new means that not only enables the acquisition of information on the base sequences of 5'-ends in mRNAs within a sample, but also enables the



cloning of new genes and the analysis of genomic sequence information, which correspond to coding and regulatory regions.

【0 0 0 8】

This can include statistics on the DNA transcriptional starting site. By using concatamers to obtain information on a large number of 5'-sequence tags as presented in the invention, it is possible to effectively map transcriptional start sites and the related promoter sequences. Thus the invention provides new means, where SAGE did not allow any promoter analysis due to the use of unrelated 3'-ends. At the same time, there were techniques for the collection of full-length cDNA clones and sequences derived thereof; however, those are focusing on collecting the full-length cDNA clones and not fragments covering the 5'-ends. Therefore full-length cDNA cloning approaches are not suitable for high throughput identification and analysis of start sites of transcription and the related promoter regions. The invention offers here a novel way to combine contrasting teachings and to obtain by a high throughput approach 5' ends, which are useful for promoter mapping and analysis. The use of the invention to study and analyze complex regulatory networks in combination with the ability to identify and clone new genes opens a wide area of applications for the invention to monitor biological systems and their statuses in development, homeostasis, and disease.

【0 0 0 9】

【課題を解決するための手段】

After devoted research, the inventors involved in this application were able to complete the present invention by arriving at the fact that by selectively collecting multiple nucleic acid fragments containing information on the base sequences at the 5' end site of the mRNAs, it is not only possible to acquire information on the base sequences in mRNAs, but it is also possible to clone new genes; and they were also able to arri

ve at a concrete method for attaining this goal.

【0010】

That is, the present invention provides a method for preparing concatemers of a plurality of nucleic acid fragments having information on nucleotide sequences of 5'-end regions of a plurality of mRNAs in a sample, comprising a first step of selectively collecting a plurality of first-strand cDNAs containing regions complementary to 5'-end regions of mRNAs, which cDNAs are formed by using mRNAs in the sample as templates; a second step of selectively collecting fragments containing cDNA regions including at least the regions complementary to the 5'-end regions of said mRNAs; and a third step of ligating the collected fragments to form a concatemer. The present invention also provides a method for determining nucleotide sequences of 5'-end regions of a plurality of mRNAs by sequencing said concatemer prepared by the method according to the present invention. The present invention further provides a method, which is the same method according to any one of claim 1 to 10, except that preliminarily obtained cDNAs having complete length is used instead of carrying out said first step. The present invention still further provides the concatemer prepared by the method according to the present invention. The present invention still further provides a vector comprising said concatemer according to the present invention. The present invention still further provides sequence tags derived from said concatemers prepared according to the present invention. The present invention still further provides means to use the sequences derived from said concatemers to analyze the content of the plurality of a RNA sample. The present invention still further provides means to use the sequences derived from said concatemers to identify regions in the genome, which are required for gene regulation and gene expression.

【0011】

The invention is not limited to the use of concatamers for sequencing of 5' ends, but modifications at particular steps of the enrichment of 5' ends and their cloning as disclosed here allow for the individual sequencing of specific 5' ends. Such embodiments of the invention would include a modification of the first and second step, where a linker would be used that is specifically bound to a solid matrix. The cDNA bound to the support would then be used to prepare the sequencing reactions.

#### 【 0 0 1 2 】

Thus the invention refers more generally to the concept of isolating portions of nucleic acids corresponding to the 5' end of transcribed genes and using them to further high-throughput analysis such as sequencing.

#### 【 0 0 1 3 】

##### 【発明の実施の形態】

As described above, the method of the present invention can comprise but is not limited to roughly three steps each of which further comprises a plurality of steps. Each step will now be explained below. The concrete working examples of each step is described in detail in the later-mentioned working examples.

#### 【 0 0 1 4 】

##### Step 1

Step 1 is a step to selectively collect nucleic acids said cDNAs containing a site corresponding to the 5' end site of mRNAs within a sample and which are synthesized for instance by using said mRNAs as templates.

#### 【 0 0 1 5 】

Either total RNA or mRNA taken from a desired cell or tissue can be used as the starting substrate. The preparation method of total RNA and mRNA is already known, and it is also described in detail in the later-mentioned working examples. In other embodiments, a full-length cDNA library

y may be used to isolate the 5' end nucleic acids corresponding to the 5' end of the transcribed part of the genes. Alternatively, a cDNA library itself would be cleaved if it carries a Class II's enzyme in proximity of the 5' end.

【 0 0 1 6 】

Step 1 itself can be conducted by a publicly known method. In other words, methods to construct full-length cDNAs and methods to synthesize cDNA fragments at least containing a site corresponding to the 5' end site of the mRNAs are already known, and any of these methods can be adopted. One of the preferable methods is the cap trapper method (e.g. Piero CARINCI et al., METHODS IN ENZYMOLOGY, VOL. 303, pp. 19-44, 1999). This cap trapper method shall be explained below, however, the invention is not limited to the use of the cap trapper method and other approaches to enrich or select full-length cDNAs could be applied as well. An alternative method (as described by Pelletier et al. in 1995) makes use of an immobilized cap-binding protein to isolate full-length cDNAs after RNase treatment of a hybrid.

【 0 0 1 7 】

Alternatively to the cap-selection, one could dephosphorylate with a phosphatase, such as BAP (bacterial alkaline phosphatase) the 5' ends of mRNAs, followed by treatment with the decapping enzyme TAP (tobacco acid pyrophosphatase). Subsequently a ribonucleotide or a deoxyribonucleotide can be attached to the 5' end of the mRNA instead of the original cap-structure with RNA ligase (Maruyama and Sugano). In this way, for instance, a Class II S recognition site could be placed on the oligonucleotide/ribonucleotide sequence using during the ligation step, which is placed at the 5' end of a cDNA or RNA. This class II's restriction enzyme can then cleave the cDNA and produce the 5' end tag.

【 0 0 1 8 】

Alternatively to biotin, a cap-binding protein (Pelletier et al. Mol Cell Biol 1995 15:3363-71) or an antibody that specifically binds to the cap structure can be used as the aforementioned selectively binding substance.

【0019】

Alternatively, one could use methods to attach oligonucleotides chemically to the cap structure as described by Genset. This method is based on the oxidation of cap (US patent 6,022,715). This allows (1) adding to the cap an oligonucleotides, which may contain the ClassII's enzyme, (2) preparing first-strand cDNA synthesis which then switched second strand cDNA synthesis; after the second strand synthesis, the cDNA would be cleaved with Class II's enzymes to make a 5' tag, for subsequent formation of the concatamer.

【0020】

Alternatively, one could use the Use the cap-switch method as described by Clontech (US patent 5,962,272). One could prepare the first-strand cDNA in presence of a cap-switch oligonucleotide, which carries a recognition site for a substance capable to recognize nucleic acids and cleave them apart from the said recognition sequence such a Class II's restriction enzyme site. The cap switch mechanism let the first strand sequence continue on the cap-switch oligonucleotides. This can be followed by second cDNA strand, possibly also followed by PCR (as describes for instance in the SMART™ Clontech cloning system), and finally it would be cleaved with the class II's to produce the 5' end TAGS.

【0021】

In another embodiment, when the quality of RNA allows it, one can prepare the cDNA by priming and extending the RNA until the cap-structure. Particular enzyme and reaction condition allow sometimes reaching the cap-site very efficiently (Carninci et al, Biotechniques, 2002). Even witho

ut a cap-selection it is possible to attach oligonucleotides in place of the cap structure, which carry Class II's restriction enzyme sites that would be later used to produce concatamers.

#### 【 0 0 2 2 】

The cap trapper method first synthesizes the first-strand cDNA with a reverse transcriptase by using RNA as a template. This can be conducted by a known method. The cDNA can be primed with an oligo-dT primer or, when the template RNA is mRNA, it can be primed with a random primer. It is advisable to add trehalose to the reactive solution because it raises the efficiency of reverse transcription reaction by stabilizing the reverse transcriptase. It is preferable to use 5-methyl-dCTP instead of standard dCTP, because it avoids internal cDNA cleavage with several restriction enzymes and prevents unintended cleavage with restriction enzymes to a considerable extent. In addition, after the first-strand cDNA synthesis, proteins and digested peptides might be removed by CTAB (cetyl trimethyl ammonium bromide) treatment, or other more general methods to purify cDNA.

#### 【 0 0 2 3 】

Next, a selectively binding substance is bound to the cap structure of mRNA. A "selectively binding substance" here means a substance that selectively binds to a specific substance, preferably but not limited to biotin. The cap structure is the structure at the 5' end of mRNA, which does not exist in transfer RNA (tRNA) or ribosomal RNA (rRNA). Therefore, even if total RNA was used as the starting substrate, the selectively binding substance only binds to mRNA. In addition, the selectively binding substance does not bind to mRNA if the cap structure at the 5' end has been cleaved. Biotin can be bound to the cap structure by a known method. For instance, the cap structure can be biotinylated by first oxidizing the diol groups on the cap structure by treating mRNA with an oxidizer

such as  $\text{NaIO}_4$  and making then react with biotin hydrazide. Alternatively, any other methods known to a person trained in the state of the art of the preparation of full-length cDNAs can be utilized to selectively enrich 5'-ends according to the invention.

【 0 0 2 4 】

Then, single-strand RNA is cleaved by means such as RNase I treatment. Any other RNase that can cleave single strand RNA but not cDNA/RNA hybrids or cocktails of RNases that can cleave the various single-strand RNA sequences at various specificity can be used alternatively. In an RNA/cDNA hybrid whose first-strand cDNA has not extended to the site corresponding to the 5' end site of RNA, the vicinity of the 5' end of RNA is single-stranded due to its failure to be hybridized with cDNA. Thus, the hybrid is cleaved at the single-stranded part and loses its cap structure through this step. Consequently, this step leaves only those mRNA/cDNA hybrids with cDNA that fully extends to the 5' end of mRNA to maintain the cap structure.

【 0 0 2 5 】

A matching selectively binding substance fixed to a support, which selectively binds to the aforementioned selectively binding substance, is prepared. In the present specification, a "matching selectively binding substance" means a substance that selectively binds to the aforementioned selectively binding substance, which, in the case where the selectively binding substance is biotin, would be avidin, streptavidin or a derivative thereof that binds specifically to biotin or its derivatives. The support can favorably be, but is not limited to be, magnetic beads, particularly magnetic porous glass beads. Since magnetic porous glass beads to which streptavidin has been fixed are commercially available, such commercial streptavidin-fixed magnetic porous glass beads can be used favorably. Similarly other materials such as latex beads, latex magnetic bead

s, agarose beads, polystyrene beads, sepharose beads or alike could be used instead of porous glass beads. Furthermore, the invention is not limited to the use the biotion-avidin system but other binding substances could be used like a digoxigenin tag that would be attached to the cap structure and digoxigenin recognizing antibodies attached to a solid matrix.

#### 【0 0 2 6】

Following this, the aforementioned mRNA/cDNA hybrid with the cap structure is made to react with the aforementioned matching selectively binding substance fixed to the support in order to bind the selectively binding substance on the cap structure with the matching selectively binding substance on the support, thereby immobilizing the mRNA/cDNA hybrid with the cap structure on the support. When magnetic beads are used as the support, the magnetic beads can be quickly collected by applying a magnetic force. As mentioned above, the mRNA/cDNA hybrids that have the cap structure at this stage are only those with cDNA that fully extends to the 5' end of mRNA, so cDNAs containing a site complementary to the 5' end of mRNAs are selectively collected by this step, and Step 1 is completed. Meanwhile, in order to prevent non-specific binding to the support, it is preferable to treat the support with a large excess of DNA-free tRNA for blocking such binding before conducting this reaction. Other substances that are suitable for blocking the surface are nucleic acids or derivatives, for instance total RNA or oligonucleotides; proteins, for instance bovine serum albumine; polysaccharides, for instance glycogen, dextran sulphate, heparin or other polysaccharides. Alternatively, hybrid molecules containing parts of all of the above could be used to mask non-specific binding sites.

#### 【0 0 2 7】

The above focuses on the case where Step 1 is conducted by the cap tra



pper method, but other various methods can also be used as indicated as long as they can selectively collect cDNAs containing a site complementary to the 5' end site of mRNA.

#### 【0028】

The following Step 2 selectively collects fragments containing a cDNA site that at least contains a site complementary to the 5' end site of mRNA.

#### 【0029】

First, the first-strand cDNA that has been immobilized on the support is released. It can be conducted by treating the support with alkali, such as NaOH. Alternatively to alkali, an enzymatic reaction with RNaseH (which cleaves only the RNA hybridized to DNA) could be used. The alkali treatment releases the cDNA from the mRNA/cDNA hybrid, bound to the support through the cap on the mRNA and separates the cDNA from the mRNA to only leave first-strand cDNA on its own.

#### 【0030】

Then, a linker carrying a sequence that can be recognized in a sequence-specific manner by a substance having an enzymatic activity that cleaves the recognized DNA outside the recognition sequence. An example of such substance is a Class IIs restriction enzyme.

#### 【0031】

In this embodiment, a linker that at least carries a Class IIs restriction enzyme site, and a random oligomer part at the 3' end site, is ligated to the end of this first-strand cDNA, which corresponds to the 5' end of the aforementioned mRNA (i.e. the 3' end of the cDNA). For the later cloning of the 5' end sequence tags into concatemers it is preferable but not essential to introduce a second recognition site into the linker, which should be distinct from the aforementioned recognition site used for the e.g. Class IIs restriction enzyme.

## 【 0 0 3 2 】

This can preferably be conducted as follows, by a method using a linker that carries a Class II's restriction enzyme site and a random oligomer part (SSLLM (single strand linker ligation method), Y. Shibata et al., BioTechniques, Vol. 30, No. 6, pp. 1250-1254, (2001)). The Class II's restriction enzyme is a restriction enzyme group that causes cleavage at parts other than the recognition site. An example includes but is not limited to the use of GsuI. GsuI treatment cleaves one of the strands at 16 bp downstream from the recognition site, and the other strand at 14 bp downstream from the recognition site. Another suitable example is MmeI, which cleaves respectively 20 and 18 bases apart its recognition sequence. The random oligomer part is located at the 3' end site of the linker, and though the number of bases is not particularly restricted, the recommended number is 5 to 9, or more preferably, 5 to 6. The Class II's restriction enzyme site should be located close to the aforementioned random oligomer part, so that the cleavage point comes within the cDNA, particularly relatively within the 5' side of the cDNA (i.e. the 3' side of the template mRNA). The linker should preferably be a linker for double-stranded DNA of which the aforementioned random oligomer part protrudes to the 3' side and provides the binding end. In addition, it is advisable to bind a selectively binding substance such as biotin to the linker in advance to facilitate its collection later.

## 【 0 0 3 3 】

When the aforementioned first-strand cDNA is made to react with such a linker, the random oligomer part of the linker hybridizes with the 3' end site of the first-strand cDNA (i.e. the 5' end site of the template mRNA). Next, the second-strand cDNA is synthesized by using this linker as a primer and the first-strand cDNA as a template. This step can be con

ducted by a standard method.

【 0 0 3 4 】

Then, the obtained double-strand cDNA is treated with the above Class IIs restriction enzyme. This step produces a double-strand cDNA fragment comprising a linker-derived part and a part derived from the 5' end site of the cDNA (the 5' end site of the second-strand cDNA). For instance, if GsuI were to be used as the Class IIs restriction enzyme and if there were to be a linker designed to locate the restriction site immediately upstream from the aforementioned random oligomer site, the obtained DNA fragment would include a site derived from the site on the 5' end side of the second-strand DNA (i.e. the site on the 5' end side of the mRNA) of the length of 16 bp (however, the complementary strand is 14 bp). In the case of the use of Mme I the length of the second-strand DNA fragment should increase to 20 and 18 bp respectively.

【 0 0 3 5 】

Next, such DNA fragments are selectively collected. If a selectively binding substance (e.g. biotin) had been bound to the linker as above, the collection could be conducted similarly to Step 1 by using a support to which a matching selectively binding substance (e.g. streptavidin) would be fixed. This procedure completes Step 2, which selectively collects fragments containing a cDNA site, belonging to the first-strand cDNA, which at least contains a site complementary to the 5' end site of the aforementioned mRNA.

【 0 0 3 6 】

The above explains the case where the SSLLM is used for Step 2, but Step 2 can also be carried out by any other method as long as the method can selectively collect fragments containing the 3' end site of the first-strand cDNA (the 5' end site of the template mRNA). For instance, it is possible to use exonuclease that cleaves the nucleotide in the 5'-3' di

rection at a controlled speed. The exonuclease treatment of the first-strand cDNA for a prescribed time period leaves a single-strand fragment comprising the 3' end site of the first-strand cDNA (the 5' end site of the template mRNA). It is possible to obtain only the targeted single-strand fragments by conducting treatment with a nuclease that only splits double-strand fragments. These fragments can be collected, joined with adapters and cloned.

#### 【 0 0 3 7 】

The subsequent Step 3 forms concatamers by mutually ligating the collected fragments. Since there are multiple mRNAs and the linker hybridizes with the first-strand cDNA at the random oligomer part as above, the above method can obtain fragments containing multiple cDNAs derived from multiple mRNAs within a sample. Step 3 ligates these multiple fragments and forms concatamers. The ligation of the cDNA fragments can be carried out by a standard method, using commercial ligation kits. The ligation can be securely conducted but is not limited to a method which first is introducing a second linker providing a recognition site for a restriction enzyme that is distinct from the other recognition sites used at the earlier stages, which is then ligating two fragments into di-tags, and which is further ligating such ligated di-tag fragments into concatamers. The number of ligated fragments is not restricted, practically any number above two and preferably about 30. The obtained concatamers are preferably but not limited to be amplified or cloned by a standard method.

#### 【 0 0 3 8 】

The concatamers obtained in this way each comprise a site having the same base sequence (however, uracil in RNA would be thymine in DNA) as that of the 5' end site of the multiple mRNAs within the sample. Although it also comprises a part derived from the linker or linkers, the base sequence of the linker or linkers is already known, so the part derived from

om the linker or linkers and the part derived from mRNA can be clearly distinguished by investigating the base sequence of the concatamer. Therefore, by determining the base sequence of the obtained concatamer, it is possible to find out the base sequences at the 5' end site of multiple mRNAs within the sample. The base sequences of a maximum of 16 or 20 bases at the 5' end site of each mRNA can be learned by the preferable mode of using GsuI or Mme I. Information on 16 or 20 bases would be sufficient for almost definitely identifying the mRNA statistically and to judge whether or not it is a new mRNA. In addition, by determining the base sequence of the concatamer, it is possible to learn the base sequences at the 5' end site of mRNAs for the number of above fragments included in the concatamer (preferably 20 to 30), so information on the 5' end site of multiple mRNAs can be determined efficiently. The analysis of the concatamers can be automated by the use of computer software to distinguish between sequences derived from the 5'-ends and sequences derived from a linker or the linkers.

#### 【0039】

When a new mRNA exists in a base sequence at the 5' end, the cDNA derived from the new mRNA can be obtained by conducting RT-PCR, making that site the forward primer and oligo-dT the reverse primer. It is also possible to amplify the mRNA by methods such as NASBA. Accordingly, the method of the present invention can be used for the cloning of new genes. Similarly, forward primers derived from 5'-end specific information can be used to amplify partial or full-length cDNA fragments from existing cDNA libraries.

#### 【0040】

While the above method had used mRNA or total RNA within the sample as the starting substrate, Step 1 can be omitted by using an existing full-length cDNA library. In this way, information on the base sequences of

the 5' end site of multiple cDNAs (i.e. the 5' end site of the mRNAs used as templates for said cDNAs) contained in the full-length cDNA library can be efficiently obtained similarly to the above procedure.

【0 0 4 1】

In some embodiments it could be desirable to obtain extended sequence information from the 5'-ends of transcribed regions. Such extended sequences may allow in specific cases for the identification of start sites of protein synthesis or a better mapping to genomic sequences. As described above the invention included in Step 2 the ligation of a linker to the 5' end of a cDNA. Such a linker can be modified by introducing a single-stranded overhang encompassing a sequence obtained from a concatamer to bind to and to be ligated to a specific nucleic acid fragment. After the ligation the linker can be used to enrich the DNA fragment by attaching the linker to a support from which it could be released after the enrichment. The linker can further be used as a primer to obtain extended sequence information on 5' ends.

【0 0 4 2】

By investigating the base sequences of the concatamers or extended 5'-sequences obtained by the present invention, it is not only possible to clone new genes as described above, but also possible to investigate the expression profiles of genes within the sample. Furthermore, the technology can be used for various purposes such as to map transcription start sites in the genome, to map promoter usage patterns, for the analysis of SNPs in promoter regions, for creating gene networks by combining the expression analysis with information on promoters, alternative promoter usage and the other data, and for selective collection of the promoter site within fragmented genomic DNA. To select genomic fragments containing promoter sites, a fragment containing the same base sequence as the 5' end site of mRNA could be bounded to a support e.g. by using the aforemen-

entioned Biotin system, and hybridized to fragmented genomic DNA. Hybridized genomic DNA fragments could then be separated from a mixture of genomic fragments by using e.g. streptavidin-fixed magnetic beads, and cloned under standard conditions.

【0043】

Alternatively, one could avoid to make concatamers and use selected 5' end tags by ligating a mixture of full-length cDNAs to magnetic beads carrying homogeneous sequence of oligonucleotides, followed by ligation such as in the SSLLM, second strand cDNA preparation and cleavage with a Class II restriction enzyme. The 5' end specific tag would be anchored specifically to the beads and would be used for the specific sequencing as done by Lynx therapeutics (US patents 6,352,828; 6,306,597; 6,280,935; 6,265,163; 5,695,934).

【0044】

For instance, oligonucleotides would have a "random part I", which will bind to 5' ends of cDNAs; and a code part of the oligonucleotide, which will be able to "tag" the ligation product. The oligonucleotide may be destroyed by exonuclease VII if not hybridized with a cDNA. The "decoder" oligonucleotides would be used to select out the sequence. The specific arrays of cDNAs on beads are then arrayed onto a solid surface, one per position, followed by parallel sequencing. If you look at 1 hole per 1 bead, you can make arrays of beads having specific oligonucleotides.

【0045】

By modifications as the aforementioned approaches for direct sequencing of 5' end the invention provides different means for the general analysis of 5' ends in the form of concatamers or the analysis of individual 5' ends, which were enriched by means of a 5' end specific selection.

【0046】

## 【実施例】

The present invention will now be described by way of examples thereof. It should be noted that the present invention is not restricted to the Examples. The experiments describe in the Examples can be performed by any person experienced in the state of the art of standard techniques in the field of Molecular Biology. Unless otherwise defined in the text, the technical terms, abbreviations, and solutions used in the Examples should have the same meaning as commonly understood by a person experienced to the state of the art in the field of the invention. A general description of such terms, abbreviations and solutions can be found in the common reagent section in Molecular Cloning (Sambrook and Russel, 2001). All publications mentioned herein are incorporated into this document by reference to be disclosed and to describe the methods and/or materials therein.

## 【0047】

## Example 1:

## Preparation of total RNA from tissue

In the literature a variety of different approaches for the preparation of RNA have been described, which are known to a person experienced in the state of the art. All such approaches should allow the preparation of a plurality of RNA samples derived from biological materials including tissues and cells, which are suitable for the invention. Below two such procedures are described in detail.

## Buffers and solutions:

- a) Solution D: 4M guanidinium thiocyanate, 25mM sodium citrate (pH7.0), 100mM 2-mercaptoethanol and 0.5% n-lauryl-sarcosine.
- b) RNase-free CTAB/UREA solution: 1% CTAB (Sigma), 4M UREA, 50mM Tris-HCl (PH 7.0), 1mM EDTA (pH 8.0).
- c) Water equilibrated phenol as described in Molecular Cloning



ng (Sambrook and Russel, 2001).

Phosphate-buffer saline (PBS) as described in Molecular Cloning (Sambrook and Russel, 2001)

5 M Sodium chloride

7 M Guanidium choride

Rnase free dd-water

### 【 0 0 4 8 】

Protocol for total RNA preparation

Dissect the tissue as fast as possible in a cooled dish.

Roughly evaluate the volume of tissue in a 50 ml falcon tube. The best quantity of tissue is between 0.5-1 g of tissue for 20 ml Solution D

Add 2 ml of 2M sodium acetate (pH 4.0) and 16 ml of water-equilibrated phenol.

Mix by a vortex. Add 4 ml of chloroform and shake vigorously by your hands and a vortex. Let it stay on ice for 15 min.

Centrifuge it at 6,000 rpm for 30 min at 4 °C

Transfer the upper aqueous phase to new tube by pipetting (25 ml) and recover approximately 20 ml thereof.

Precipitate the RNA from the aqueous phase by adding 1 equal volume of Isopropanol (in this case, approximately 20 ml), store on ice for 1 h.

Centrifuge at 7,500 rpm for 15 min at 4 °C: RNA is pelleted by centrifugation.

The pellet is washed twice with 70% ethanol, each time followed by centrifugation at 7,500 rpm for 2 min, in order to remove the SCN salts.

CTAB removal of polysaccharides. Selective CTAB precipitation of mRNA is performed after complete RNA re-suspension in 4 ml of water. Subsequently, 1.3 ml of 5 M NaCl is added and the RNA is then selectively precipitated by adding 16 ml of a CTAB/urea solution.

Centrifuge for 15 min at 7500 rpm (9500 x g), discard the aqueous phase.

Resuspend the RNA pellet in 4 ml of 7 M Guanidinium Chloride.

Re-suspended RNA is finally precipitated by adding 8 ml of ethanol. Incubate on  $-20^{\circ}\text{C}$  for 1-2 hours (or longer) and centrifuge for 15 min at 7,500rpm,  $4^{\circ}\text{C}$ . At the end, wash the pellet with 5 ml of 70% ethanol.

Centrifuge again at 7,500 rpm for 5 min.

Discard the supernatant.

Re-suspend RNA in 500-1000 microL of RNase-free dd-water.

#### 【 0 0 4 9 】

Preparation of a mRNA fraction from total RNA

The mRNA fraction of total RNA preparations can be isolated by the use of commercial kits such as the MACS mRNA isolation kit (Milteny) or polyA-quick (Stratagene), which provide satisfactory yield of mRNA under the recommended conditions. One cycle of oligo-dT selection of the mRNA is sufficient. It is advisable to redissolve the poly-A<sup>+</sup> RNA at a high concentration of 1 to 2 microG/microL.

#### 【 0 0 5 0 】

Preparation of a plurality of RNA samples from a cDNA library

Alternatively, a plurality of nucleic acids corresponding to the 5' ends of genes can be obtained from existing cDNA libraries, which were cloned into expression vectors. By standard methods known to a person familiar with the state of the art of molecular biology approaches, from such libraries RNA transcripts can be obtained by in vitro transcription reactions using e.g. a T3, T7 or SP6 RNA polymerase. Such an approach can be performed by first linearization of the plasmid DNA with appropriate restriction endonucleases. The restriction enzyme can be chosen to allow for the transcription of the sense RNA. In the case of libraries obtained in the vector pFLC III (Carninci P, Shibata Y, Hayatsu N, Itoh M, Shiraki T, Hirozane T, Watahiki A, Shibata K, Konno H, Muramatsu M, Hayashizaki Y.) Balanced-size and long-size cloning of full-length, cap-trapped

cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis, Genomics, 2001 Sep;77(1-2):79-90), the vector can be linearized by cleavage with one of the homing endonucleases I-Ceu I or PI-Sce I to avoid a truncation of the inserts. For the digest mix in a tube

Plasmid DNA	100 microG
10x buffer	40 microL
Restriction enzyme	100 u
ddH <sub>2</sub> O	ad 400 microL

Incubate at appropriate temperature for at least 2h and analyze 1 microL of the reaction mixture by agarose gel electrophoreses. If the digest is completed, add:

0.5 M EDTA	8 microL
10% SDS	8 microL
Proteinase K (10 mg/ml)	5 microL

Incubate for 15 min at 45°C before extracting sample with 500 microL phenol/chloroform. The aqueous phase is to be re-extracted twice with 500 microL chloroform. Finally linearized DNA is precipitated with isopropanol or ethanol under standard conditions and dissolved in 50 microL TE.

### 【 0 0 5 1 】

In vitro RNA synthesis:

Mix in a tube under Rnase free conditions:

Linearized plasmid DNA	20 microG
5x T7 or T3 buffer	200 microL
0.1 M DTT	100 microL
2 mg/ml BSA	40 microL
10 mM rNTPs	50 microL
T7 or T3 RNA polymerase	10 microL
ddH <sub>2</sub> O	ad 1000 microL

Incubate at 37°C for 3 to 4 h before adding:

10 mM Calcium Chloride 10 microL

1U/microL DNase RQ1 5 microL

Incubate at 37°C for 20 min before adding:

0.5 M EDTA 10 microL

10 mg/ml Protease K 5 microL

Incubate at 45°C for 30 min, before addition of Sodium Chloride to a final concentration of 1M. Phenol/Chloroform extraction followed by re-extraction with Chloroform should be performed under standard conditions, and the RNA transcripts can be finally collected by Isopropanol or Ethanol precipitation. The pellet is to be resuspended in 200 microL of water or TE. The quality of the RNA transcripts should be confirmed by agarose gel electrophoresis and quantification.

#### 【 0 0 5 2 】

2. : First strand cDNA synthesis

Buffers and solutions

Saturated Trehalose, about 80% in water (crystals will remain), low metal content

4.9 M high purity sorbitol

Optionally: Takara GC-Taq buffer

#### 【 0 0 5 3 】

Enzymes and buffers

RNase H<sup>-</sup> reverse transcriptase Superscript II (Invitrogen) and buffer or other reverse transcriptases.

#### 【 0 0 5 4 】

Nucleic acids and oligonucleotides

Purified, first-strand oligo-dT primer (Sequence for primer used:

5'-GAGAGAGAGAGGATCCTTCTGGAGAGTTTTTTTTTTTTTTVN-3'). Alternatively or additionally, random primer (dN<sub>6</sub>-dN<sub>9</sub>), where N is any nucleotide.

mRNA, recommended 2.5 to 25 microG or alternatively, total RNA, 5-50 microG

【 0 0 5 5 】

Radioactive compounds

[alpha-<sup>32</sup>P] dGTP

【 0 0 5 6 】

Protocol A: Trehalose-Sorbitol enhanced

To prepare the 1<sup>st</sup> strand cDNA, put together the following reagents in three different 0.5 ml PCR tubes (A, B, and C)

【 0 0 5 7 】

Tube A: in a final volume of 21.3 microL, add the following:

mRNA                    2.5-25 microG

or total RNA,                    5-50 microG

1<sup>st</sup> strand primer (2 microG/microL)                    14 microG (7 microL)

Total volume: 22 microL

Heat the mixture (mRNA, primer) at 65°C for 10 min to dissolve the secondary structures of mRNA.

Tube B: in a final volume of 76 microL, add the following:

5X 1<sup>st</sup> strand buffer                    28.6 microL

0.1 M DTT                    11 microL

dATP, dTTP, dGTP, and 5-methyl-dCTP 10 mM each                    9.3 microL

4.9 M sorbitol                    55.4 microL

Saturated trehalose                    23.2 microL

RNase H<sup>-</sup> Superscript II reverse transcriptase (200 U/microL)

15.0 microL

Final volume: 142.5 microL

【 0 0 5 8 】

Prepare a cycle (on a thermal cycle) with: 40°C, 4 min; 50°C, 2 min; 56°C, 60 min.

If total RNA is used as the starting material, prepare a cycle with:  
40°C, 2 min, -0.1°C/sec to 35°C; 50°C, 2 min; 56°C, 60 min.

Alternatively: prime the cDNA with a random primer (dNg, N= any nucleotide) at 25°C.

【 0 0 5 9 】

Tube C:

1~1.5 microL of [ $\alpha$ -<sup>32</sup>P] dGTP.

【 0 0 6 0 】

For a cold-start operate as follows:

Quickly mix tubes A and B on ice.

Transfer in tube C 40 microL of the A+B mixture.

Tubes A+B and C should be quickly transferred immediately at 40°C of the  
step 1 of the above cycling program to anneal at 40°C four 4 minutes.

Let the reaction proceed following the thermal cycler setting.

【 0 0 6 1 】

For a hot-start, operate as follows:

Transfer the tubes A, B, C on the thermal cycler

Start the cycling

When the temperature reaches 42°C, quickly mix tubes A and B.

Transfer in tube C 40 microL of the A+B mixture.

Let the reaction proceed following the thermal cycler setting.

【 0 0 6 2 】

Protocol B: GCI-Trehalose-Sorbitol enhanced

Tube A: in a final volume of 22 microL, add the following:

mRNA                      5-25 microG

(precipitate with ethanol and re-suspend directly with the primer)

or total RNA, up to 50 microG (for the small-scale protocol)

Purified 1<sup>st</sup> strand cDNA primer (2 microG/microL) 14 microG (7 microL)

Final volume: 22 microL

Tube B: add the following:

2 X GC I (LA Taq) buffer (TaKaRa)	75microL
dATP, dTTP, dGTP, and 5-methyl-dCTP, 10 mM each	4 microL
4.9 M sorbitol	20 microL
Saturated trehalose (approximately 80%)	10 microL
Superscript II reverse transcriptase (200 U/microL)	15 microL
ddH <sub>2</sub> O	4 microL
Final volume:	128 microL

Tube C:

alpha- <sup>32</sup> P-dGTP	1.5 microL
-----------------------------	------------

For the rest of the procedure, follow exactly the point as in the normal reaction condition. Prepare (in advance) a thermal cycler with the following cycle:

42°C, 30 min; 50°C, 10 min; 55°C, 10 min; 4°C, indefinite time.

【 0 0 6 3 】

Operate as follows:

- 1) Transfer the tubes A, B, C on the thermal cycler
- 2) Start the cycling
- 3) When the temperature reaches 42°C, quickly mix tubes A and B.

- 4) Transfer in tube C 40 microL of the A+B mixture.
- 5) Let the reaction proceed following the thermal cycle r setting.

At the end, stop the reaction with EDTA at 10 mM final concentration.

Then incorporation of [alpha-<sup>32</sup>P]GTP is measured and the yield of cDNA is calculated. Calculation of the amount of cDNA by measuring [alpha-<sup>32</sup>P]GTP is useful for monitoring whether the processes are accurately proceeding or not.

【 0 0 6 4 】

3. CTAB precipitation of the first-strand cDNA

Buffers and solutions

CTAB solution as described in Example 1

After measuring the radioactivity, transfer both the "hot" and "cold" 1<sup>st</sup> strand synthesis (tube B and C) to a tube and perform CTAB precipitation as follows.

Mix the tube B and C from the first strand; to the mixture add:

3 microL of 0.5 M EDTA (final concentration of 10 mM)

2 microL of 10 microG/microL Proteinase K.

Incubate at 45°C or 50°C for at least 15 min, and as long as 1 hour.

To the 128-142 microL volume of the first strand cDNA reaction, add:

32 microL of 5 M Sodium Chloride (RNase free)

320 microL of CTAB-Urea solution

Incubate at room temperature for 10 min.

Centrifuge at 15,000 rpm for 10 min

Remove supernatant.

Carefully re-suspend with 100 microL of 7M Guanidinium Chloride

Add 250 microL of ethanol and leave on ice or 20 to -80°C for 30-60 min

Centrifuge at 15,000 for 10 min. Remove the supernatant.

Subsequently, wash the pellet twice with 800 microL of 80% ethanol. Each time, add 80% ethanol to the tube and centrifuge for 3 min. at 15,000 rpm.

Re-suspend cDNA in water 46 microL.

【 0 0 6 5 】

4. Cap-trapping, oxidation and biotinylation of the cap

Buffers and solutions

1 M sodium acetate buffer, pH 4.5

1M citrate buffer, pH 6.0



NaIO<sub>4</sub>, solution >100 mM.

SDS 10%

Biotinylation buffer: 33 mM Sodium citrate, pH 6.0, and 0.33% SDS.

10 mM Biotin Hydrazide long arm (MW = 371.51; 3.71 mg/ml = 10 mM) in citrate/SDS buffer.

Cap biotinylation: (A) Oxidation of the diol groups of mRNA

In a final volume of 50 to 55 microL, add the following:

The re-suspended cDNA sample

3.3 microL of 1 M sodium acetate buffer, pH 4.5

A freshly prepared solution of NaIO<sub>4</sub> to a final concentration of 10 mM

Incubate on ice in the dark for 45 min.

Finally, precipitate the cDNA:

To simplify the downstream process, add 1 microL of glycerol 80%.

Vortex.

Add 0.5 microL of 10% SDS, 11 microL of 5 M sodium chloride and 61 microL of isopropanol.

Incubate at 20 or -80°C for 30 min in the dark.

Centrifuge for 15 min at 15,000 rpm.

Remove supernatant.

Add 500 microL of 80% ethanol

Centrifuge at 15,000 rpm for 2-3 min.

Discard the supernatant

Repeat steps 12-13

Re-suspend the cDNA in 50 microL of water.

Biotinylation: (B) Derivatization of the oxidized diol groups

To the cDNA (50 microL), add 160 microL of the dissolved biotin hydrazide long arm in the reaction buffer. Perform the reaction in 210 microL (final volume).

Incubate overnight (10-16 hours) at room temperature (22-26°C).

Subsequently, to precipitate the biotinylated cDNA, add:

75 microL 1 M Sodium citrate, pH 6.1

5 microL of 5 M Sodium chloride

750 microL of absolute ethanol

Incubate on ice for 1 hour or at 80 or -20°C for 30 min or longer.

Centrifuge the sample at 15,000 rpm for 10 min

Wash the precipitate twice with 70% or 80% ethanol and centrifuge.

Discard the supernatant and repeat the wash. Dissolve the cDNA in 175 microL of TE (1 mM Tris, pH 7.5, 0.1 mM EDTA).

Cap-trapping and releasing the 5' ends of cDNA Enzymes and buffers

RNase ONE (Promega) and its reaction buffer

To the cDNA sample add, in a final volume of 200 microL:

20 microL of RNase I buffer (Promega).

1 units of RNase I (Promega, 5 or 10 U/microL) per each 1 microG of starting mRNA or total RNA (in case of small scale protocol) used for first strand cDNA synthesis.

Incubate at 37°C for 30 min.

To stop the reaction, put the sample on ice and add

4 microL 10% SDS and

3 microL of 10 microG/microL Proteinase K.

Incubate at 45°C for 15 min.

Extract once with 1:1 Tris-equilibrated phenol:chloroform, then load the aqueous phase into Microcon -100.

Perform a back extraction with water and load again into the Microcon-Centricon 100 filter.

Perform one round of Microcon separation

8-b) Dissolve completely the pellet with 20 microL of 0.1 x TE

## 【 0 0 6 6 】

Magnetic beads blocking

Materials

Streptavidin-coated MPG (CPG inc., New Jersey)

Buffers and solutions

Binding buffer: 4.5 M NaCl, 50 mM EDTA, pH 8.0

Special equipments

A magnetic stand to hold 1.5 ml tubes is required.

To further minimize the non-specific binding of nucleic acids, magnetic beads are pre-incubated with DNA-free tRNA (10mg/ml).

For each preparation, pre-incubate 500 microL of magnetic beads (per 25 microG of starting mRNA) with 100 microG of tRNA.

Incubate on ice for 30 min with occasional mixing.

Separate the beads with a magnetic stand (for 3 min) and remove the supernatant.

Wash for 3 times with 500 microL of binding buffer

## 【 0 0 6 7 】

5'-ends cDNA capture and release

To capture the full-length cDNA, mix the RNaseI-treated cDNA and wash beads as follows:

- 1) Re-suspend the beads in 500 microL of wash/binding buffer.
- 2) Transfer 350 microL of the beads into the tube containing the biotinylated first-strand cDNA.
- 3) After mixing gently rotate the tube for 10 min at 50

0°C,

4) Transfer 150 microL of the beads into the tube containing the biotinylated first-strand cDNA and 350 microL of beads.

5) After mixing gently rotate the tube for 20 min at 50 °C.

Separate the beads from the supernatant on a magnetic stand.

Washing the beads

Gently wash the beads with 0.5 ml of the indicated buffer to remove the nonspecifically absorbed cDNAs.

2 x with washing/binding solution.

1 x with 0.3 M NaCl/ 1mM EDTA

2 x with 0.4% SDS/ 0.5 M NaOAc/ 20 mM Tris-HCl pH 8.5/ 1mM EDTA.

2 x with 0.5 M NaOAc/ 10 mM Tris-HCl pH 8.5/ 1mM EDTA.

Alkali release (see below)

Alkali full-length cDNA release from beads

Add 100 microL of 50 mM NaOH, 5 mM EDTA.

Briefly stir and incubate 5 min at RT with occasional mixing.

Separate the magnetic beads and transfer the eluted cDNA on ice.

Repeat the elution cycle with 100 microL of 50 mM NaOH, 5 mM EDTA, two more times until most of the cDNA, 80-90% as measured by monitoring the radioactivity, can be recovered from the beads.

Adding a 5'-end primable site to the cDNA

RNase step

Enzymes and buffers

- RNase ONE<sup>TM</sup> and its buffer (Promega)

Add 50 microL of 1 M Tris-HCl, pH 7.0 in tubes on ice and mix quickly.

Add 1 microL of RNase I (10U/microL) and mix quickly.

Incubate at 37 °C for 10 min.

To remove the RNaseI, treat the cDNA with Proteinase K and phenol/chloro

form extraction including back extraction.

Add 3 microG of glycogen. Treat the cDNA with one cycle of Microcon-100.

Fractionation of cDNA before adding a primable site

#### Materials

Amersham-Pharmacia S-400 spun kit or alternative kits

#### Buffers and solutions

Column buffer: 10 mM Tris, pH 8.0, 1 mM EDTA, 0.1 % SDS, and 100 mM NaCl

Column buffer without SDS: 10 mM Tris, pH 8.0, 1 mM EDTA and 100 mM NaCl

#### S-400 spun column chromatography

Detailed protocols are described in the kits. This is the running protocol of S-400 spun columns.

#### Shake the column

Break the seal and transfer in a 2 ml tube

Centrifuge at 3,000 rpm 1 min (+ 4 °C)

Add the cDNA (< 20 microL volume)

After cDNA, add 80 microL of water

Centrifuge 2 min at 3000 rpm

Concentrate by Microcon 100 or precipitate with isopropanol. Recovery should exceed 80%.

【 0 0 6 8 】

#### 6. SSLM

#### Materials

S-300 spun column chromatography kit (Amersham-Pharmacia)

#### Buffers and solutions

Column buffer: 10mM TrisHCl pH 8.0, 1mM EDTA, 0.1% SDS, 100mM NaCl.

#### Enzymes and buffers

Takara DNA Ligase KIT II.

Nucleic acids and oligonucleotides

In the Example given here, the recognition sites for the restriction enzymes Bgl II, Gsu I and Mme I are introduced, however, the invention is not dependent or limited to the use of those restriction enzymes and their recognition sites. In particular, Bgl II (recognition site: AGATCT) can be replaced by any endonuclease suitable for cloning. Other example for such enzyme could include Asc I (recognition site: GGCGCGCC) or Xba I (recognition site: TCTAGA).

- Synthesize the following oligonucleotides containing the GsuI restriction site.

Oligonucleotide Bg-Gsu-GN5:

5' -Biotin-AGAGAGAGAACTAGGCTTAATAGGTGACTAGATCTGGAGNNNNN-3' ;

Oligonucleotide Bg-Gsu-N6:

5' -Biotin-AGAGAGAGAACTAGGCTTAATAGGTGACTAGATCTGGAGNNNNNN-3' ;

Oligonucleotide Bg-Gsu-down:

5' P-CTGGAGATCTAGTCACCTATTAAGCCTAGTTCTCTCT-NH<sub>2</sub> 3'.

- Synthesize the following oligonucleotides containing the Mme I restriction site.

Oligonucleotide Bg-Mme-GN5:

5' -Biotin-AGAGAGAGAACTAGGCTTAATAGGTGACTAGATCTTCCRACGNNNNN-3' ;

Oligonucleotide Bg-Mme-N6:

5' -Biotin-AGAGAGAGAACTAGGCTTAATAGGTGACTAGATCTTCCRACNNNNNN-3' ; Oligonucleotide Bg-Mme-down:

5' P-GTYGGAGATCTAGTCACCTATTAAGCCTAGTTCTCTCT-NH<sub>2</sub> 3' .

Where R stands for G or A and Y stands for C or T.

P means that the oligonucleotide must be 5' phosphorylated and NH<sub>2</sub> indicates that an amino-group is added to avoid non-specific ligation and possible hairpin priming.

Oligonucleotides should be purified by acrylamide gel electrophoresis following standard techniques as the first-strand cDNA primer with 10% ac

ylamide electrophoresis (Sambrook and Russel, 2001). Oligonulceotides should be extracted with phenol/chloroform, chloroform and precipitation with 2 volumes of ethanol as for the first-strand cDNA primer.

#### 【 0 0 6 9 】

Preparation of the linkers.

After OD checking and mixing Bg-Gsu-GN5, Bg-Gsu-N6 and "down" oligonucleotides at ratio 4:1:5, at least 2 microG/microL of DNA; add NaCl at 100 mM final concentration. The oligonulceotides are annealed at 65°C for 5min, 45°C for 5min, 37°C for 10min, 25°C for 10min.

#### 【 0 0 7 0 】

Ligation of the first-strand cDNA

Use 2 microG of linker mixture for up to 1 microG single-strand cDNA. Mix linkers and cDNA (final volume: 5 microL)

Heat at 65°C for 5min to melt secondary structures of single-strand cDNA  
Transfer the linker and cDNA mix on ice.

Add 5 microL of the solution II from the TAKARA DNA ligation Kit.

Add 10 microL of solution I of the kit.

Incubate at 10°C overnight (at least >10 hours).

At the end of the ligation reaction, stop the reaction by adding 1microL of 0.5 M EDTA, 1 microL of 10% SDS, 1microL of 10 mg/ml Proteinase K, 10 microL of water, and incubate at 45°C for 15 min.

Treat with phenol/chloroform, chloroform and back extract (see appendix)  
with 60 microL of column buffer

After the ligation, remove the excess linker with S-300 spin column chromatography

- 1) Shake the column several times and then let it stand upright.
- 2) Remove the upper cap, then the bottom one.
- 3) Drain the buffer of the column. Apply 2 ml of the column buffer and drain twice by gravity.

Put the column into a 15 ml centrifuge tube, then centrifuge at 400 x g for 2 min in a swing-out rotor at room temperature.

Apply 100 microL of buffer to the column, then centrifuge at 400 x g for 2 min. Check the eluted volume. If it is different from the input (100 microL), repeat this step until the eluted volume is the same as the added one.

Set a 1.5 ml tube, after cutting off the cap, into the 15 ml centrifuge tube, and then apply the sample into the column. Centrifuge at 400 x g for 2 min.

Collect the eluted fraction in a separate tube. Apply to the column 50 microL of buffer, repeat the centrifugation and collect the fraction in a separate tube.

Repeat step 6 for 3 to 5 more times; keep the eluted fractions separate. Collected fractions should be counted in a scintillation counter. Usually mix the first 2-3 fractions (80% of cpm of cDNA).

Add NaCl to a final concentration of 0.2 M, precipitated the cDNA by adding equivalent of isopropanol.

After precipitation and washing twice with 80% cold ethanol, re-suspend with water.

#### Second-strand cDNA

Setting the 2nd strand cDNA program on the thermal cycler as follows:

Step 1	5 min at 65 °C
Step 2	30 min at 68 °C
Step 3	72 °C for 10 min
Step 4	+4°C

#### Procedure for the second strand cDNA

Second strand steps, mix in a test tube:

The cDNA



6 microL of LA-Taq polymerase buffer (Takara)

6 microL of 2.5 mM (each) dNTP's (Takara)

0.5 microL of [ $\alpha$ - $^{32}$ P] dGTP (optional to follow the incorporation)

After starting the 2nd strand program, put the tube on the thermal cycle  
r.

Add to tube 3 microL of 5 U/microL of LA Polymerase or alternative therm  
ostable polymerase cocktails, when the samples are at 65°C, during the fi  
rst step.

Mix quickly but thoroughly

At the end of the cycle of the thermal cycler, stop the reaction by addy  
ing 10 mM EDTA (final concentration) and clean up the reaction by Protei  
nase K treatment, Phenol-chloroform extraction and ethanol precipitation  
(see Sambrook and Russel, 2001, Molecular Cloning, CSHL press, NY).

### 【 0 0 7 1 】

#### 11. Cleavage of cDNA

The cDNA should then be cleaved with the Class IIs restriction enzyme li  
ke Gsu I given in this Example.

Buffer (10X) (MBI Fermentas)	10 microL
------------------------------	-----------

GsuI(1U/microL) (use 5U/microG DNA)	Y microL
-------------------------------------	----------

ddH <sub>2</sub> O	X microL
--------------------	----------

Final volume	100 microL
--------------	------------

Where the Y and X vary depending on the quantity of cDNA

1) Incubate at 37°C for 1 hour.

2) Added 0.5M EDTA 2 microL.

3) Incubated at 65°C for 15 min. to inactivate the enzyme

Prepare the magnetic beads

Prepare the appropriate quantity of CPG-MPG (Magnetic porous glass beads

). The same considerations made for the cap-trapper step are valid at this point.

Prepare 200 microL of GPG- beads.

Add 5 microG of tRNA (20 mg/ml).

Incubate at RT for 10-20 min or on ice for 30-60 min, with occasional shaking

Transfer the beads on a magnetic stand for 3 minutes and remove the aqueous phase.

Wash 3 times with: 1M NaCl, 10 mM EDTA use at least a volume equivalent to the starting volume of beads.

Re-suspend beads in 1M NaCl, 10 mM EDTA equivalent to the starting volume of beads.

#### 【 0 0 7 2 】

7. Released cDNA tags

Mixed washed beads and GsuI cut sample.

Incubate at RT for 15 min with occasional gentle mixing

Let it stand on magnetic rack for 3 min.

Recover the supernatant.

Rinse 4X with 500 microL of 1X B&W buffer (binding and washing buffer= 5 mM Tris, pH 7.5, 0.5 mM EDTA, and 1 M NaCl) containing 1X BSA (bovine serum albumin) wash.

Wash 2X with 200 microL of 1X ligase buffer (NEB).

#### 【 0 0 7 3 】

8. Ligating linkers to bound cDNA: II linker ligation.

In this Example a linker with a recognition site for the restriction enzyme Eco RI is used. However, the invention is not dependent or limited to the use of Eco RI in the second linker. Any other restriction enzyme and its recognition site can be used depending on their convenience for cloning the concatamers.

Oligonucleotides to be synthesized:

5'-GAGAGAGAGACTTTAGGTGACACTATAGAAGAGTCCTGAGAATTCNN-3'

5'-P-GAATTCTCAGGACTCTTCTATAGTGTCACCTAAAGTCTCTCTCTC-3'

The oligonucleotides are purified and annealed as described for the Linker 1.

LoTE (1 mM Tris, pH 7.5, and 0.1 mM EDTA) 20 microL suspended and add linker II (0.4 microG/microL)

Heat the tube at 65 °C for 5min, then let sit at room temperature for 15 min.

Add TaKaRa ligation kit II solution II 25microL and solution I 50microL. Incubated at 16 °C overnight.

After ligation, wash 4 times with 500 microL 1X B&W buffer containing 1X BSA.

Wash once with 200 microL 1X B&W buffer and twice with 200 microL 1XBglI I buffer containing 1X BSA.

#### 【 0 0 7 4 】

Release of cDNA tags using the Tagging Enzyme

Add to the sample the following

- LoTE X microL
- 10X buffer 10 microL
- Bgl II Y microL

Make up the volume to a total of 100 microL.

- 1) Incubate at 37°C for 1 hour, gently mixing intermittently.
- 2) Place on magnet, collect supernatant into new tube. The supernatant contains the released 5' end fragments.

3) Raise volume to 200 microL with LoTE.

To 200 microL of sample (the 5' ends, tagged with linkers) add:

133 microL 7.5M NH<sub>4</sub>OAc

3 microL 1microG/microL glycogen

340 microL Isopropanol

Incubate at 20 or -80°C for at least 30 min.

Spin for 20min at 4°C at 15,000 rpm in a micro-centrifuge. Remove the supernatant. Wash the pellet twice with 80% or 70% ethanol. Centrifuge for

3 min at 15,000 rpm and removed the ethanol wash. At the end, re-suspended in 10 microL LoTE.

### 【 0 0 7 5 】

Ligating tags to form di-tags

The 5' ends of cDNAs are ligated to form di-tags.

1) Add the TaKaRa ligation Kit II solution II 10 microL and solution I 20 microL.

2) Incubate overnight 16°C.

3) Added 10 microL of ddH<sub>2</sub>O, 1 microL of 0.5M EDTA, microL of 10% SDS 1 and 1 microL of 10 microG/microL Proteinase K.

4) Incubate at 45°C for 15min.

5) Extract once with 1:1 Tris-equilibrated phenol:chloroform aqueous phase. After phenol-chloroform and chloroform, and back extraction.

6) Removal the smallest cDNA fragment with a G-50 spun-column (Size exclusion).

7) precipitate with isopropanol by adding 5 microG of glycogen as carrier.

100 microL sample

67 microL 7.5M NH<sub>4</sub>OAc

5 microL glycogen

180 microL Isopropanol

- 8) Spin for 20 min at 4 °C.
- 9) Wash twice with 80% or 70% ethanol, centrifuge and remove the ethanol.

【 0 0 7 6 】

12. Cleavage of cDNA with anchoring enzyme

- 1) Re-suspend the sample in 5 microL of LoTE. Add then in order:

LoTE	X microL
10X EcoRI restriction buffer	5 microL
EcoRI	Y microL (use 20 Units of EcoRI)

Bring up the volume to a total of 50 microL.

- 2) Incubate at 37°C for 1hour.
- 3) Add 1 microL of 0.5M EDTA, 1microL of 10% SDS 1 and 1 microL of 10 microG/microL Proteinase K 10%.
- 4) Incubate at 45 °C for 15min.
- 5) Extract once with 1:1 Tris-equilibrated phenol:chloroform aqueous phase. After phenol-chloroform and chloroform, and back extraction
- 6) precipitate with isopropanol by adding 5 microG of glycogen as carrier.

100 microL sample

67 microL 7.5M NH<sub>4</sub>OAc

5 microL glycogen

180 microL Isopropanol

- 8) Spin for 20 min at 4°C.
- 9) Wash twice with 80% or 70% ethanol, centrifuge and removed the ethanol wash each time.

【 0 0 7 7 】

11. Ligation of di-tags to form concatemers
  - 1) Resuspended LoTE 5 microL.
  - 2) Added TaKaRa ligation kit II solution II 5 microL and solution II 10 microL.
  - 3) Incubate 1.5 hours at 16 °C.
  - 4) Added 0.5M EDTA 1 microL, 10% SDS 1 microL, 10 microG/microL Proteinase K 1 microL.
  - 5) Incubate at 45°C for 15min.
  - 6) Extract once with 1:1 Tris-equilibrated phenol:chloroform aqueous phase. After phenol-chloroform and chloroform, and back extraction.
  - 7) precipitate with isopropanol by adding 5 microG of glycogen as carrier.  
100 microL sample  
67 microL 7.5M NH<sub>4</sub>OAc  
5 microL glycogen  
180 microL Isopropanol
  - 8) Spin for 20min at 4°C.
  - 9) Wash twice with 80% or 70% ethanol, centrifuge and removed.  
Resolved 5 microL ddH<sub>2</sub>O.

### 【 0 0 7 8 】

#### Example 2:

The above-obtained concatamers are to be further ligated into a cloning vector such as pBlueascript II KS+ (Stratagene). A large variety of cloning vectors are known in the filed, which can be use for invention.

#### Standard Ligation:

Mix a three time excess of concatamer DNA and 100 ng of an appropriate vector linearized with Eco RI in a volume of 5 microL. Then mix 5 microL

of Solution I of DNA Ligation Kit Ver.2 (Takara) to the insert/vector mixture. Incubate the tube at 16°C for 12-16 h.

#### 【 0 0 7 9 】

##### Transformation:

To remove salt from the ligation solution, precipitate DNA after the addition of 2 microG of Glycogen (Roche), 20mM Sodium Chloride and 80% ethanol. The DNA pellet is washed twice with 150 microL of 80% of ethanol, and the pellet is then dissolved in 10 microL of water. Using 1 microL of desalted ligation solution, ElectroMAX™ DH10B™ Cells (Invitrogen) are re transformed using Cell-Porator or alike (Biometrer according to the transformation procedures described in the manufacturer's manual. Transformed bacteria are plated on a selective medium and grown overnight. Positive clones are to be isolated from those plates for further characterization of the concatamers.

#### 【 0 0 8 0 】

##### Example 3: Sequencing of concatamers

Sequencing of concatamers is performed using primers nested in the flanking regions of the cloning vector and a BigDye Terminator Cycle Sequencing Ready Reaction Kit v2.0 (Applied Biosystems) and an ABI3700 (Applied Biosystems) sequencer according to the manufacture's product descriptions. The concatamers are sequenced from both ends to cover their entire sequence.

#### 【 0 0 8 1 】

##### Example 4: Identification of 5'-end sequence tags

The sequences obtained form Concatamers are characterized by the structure of the di-tags as presented in Figure 5. Defined regions holding the recognition sites for the restriction enzymes used during the cloning steps flank each 5' end specific sequence tag. Therefore the 5' end specific sequence tags can be identified by a manual sequence analysis or by

an automated process using an appropriate computer program. Individual 5' end specific sequence tags can be stored in a computer file or a data base system.

【 0 0 8 2 】

#### Example 5: Characterization of 5'-end sequence tags

5' end specific sequence tags can be analyzed for their identity by standard software solutions to perform sequence alignments like NCBI BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>), FASTA, available in the Genetics Computer Group (GCG) package from Accelrys Inc. (<http://www.accelrys.com/>) or alike. Such software solutions allow for an alignment of 5' end specific sequence tags among one another to identify unique or non-redundant tags, which can be further used in

Database searches

Building a 5'-end sequence database

Gene identification using a 5'-end sequence database

An example of a BLAST search in GenBank using a 5' end specific tag is given below: The 16 bp tag (5'-ACC TCC CTC CGC GGA G) is derived from the 5' end of Human TGF- $\beta$ 1: JBC 264 (1989) 402-408.

Query= (16 letters) (ACCTCCCTCCGCGGAG)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)

1,205,903 sequences; 5,297,768,116 total letters

Score E

Sequences producing significant alignments: (bits)

Value

gi|10863872|ref|NM\_000660.1| Homo sapiens transforming grow... 32

1.1



gi 18590091 ref XM_085882.1  Homo sapiens similar to transf...	32
1.1	
gi 11424057 ref XM_008912.1  Homo sapiens transforming grow...	32
1.1	
gi 7684381 gb AC011462.4 AC011462 Homo sapiens chromosome 1...	32
1.1	
gi 15027087 emb AL389894.4 LMFLCHR4A Leishmania major Fried...	32
1.1	
gi 1943914 gb U70540.1 LMU70540 Leishmania mexicana amazone...	32
1.1	
gi 37097 emb X05839.1 HSTGFBG1 Human transforming growth fa...	32
1.1	
gi 37092 emb X02812.1 HSTGFB1 Human mRNA for transforming g...	32
1.1	
gi 340526 gb J04431.1 HUMTGFB1PR Homo sapiens transforming ...	32
1.1	
gi 18858696 ref NM_131728.1  Danio rerio forkhead box Cla (...)	30
4.2	
gi 12004937 gb AF219949.1 AF219949 Danio rerio forkhead tra...	30
4.2	
gi 193604 gb M13366.1 MUSGPDX Mouse glycerophosphate dehydr...	30
4.2	
gi 193601 gb M25558.1 MUSGPD Mouse glycerol-3-phosphate deh...	30
4.2	
gi 63465 emb V00414.1 GGHI01 Gallus gallus mRNA coding for ...	30
4.2	
gi 63444 emb X13894.1 GGH2AF Chicken histone H2A.F gene	30
4.2	

## Alignments

>gi|10863872|ref|NM\_000660.1| Homo sapiens transforming growth factor, beta 1

(Camurati-Engelmann disease) (TGFB1), mRNA

Length = 2745

Score = 32.2 bits (16), Expect = 1.1

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 1 acctccctccgcggag 16

>gi|18590091|ref|XM\_085882.1| Homo sapiens similar to transforming growth factor, beta 1 (H.

sapiens) (LOC147760), mRNA

Length = 697

Score = 32.2 bits (16), Expect = 1.1

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 7 acctccctccgcggag 22

>gi|11424057|ref|XM\_008912.1| Homo sapiens transforming growth factor, beta 1 (TGFB1), mRNA

Length = 2741

Score = 32.2 bits (16), Expect = 1.1

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 1 acctccctccgcggag 16

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,  
or phase 0, 1 or 2 HTGS sequences)

Posted date: Apr 9, 2002 10:59 AM

Number of letters in database: 1,002,800,820

Number of sequences in database: 1,205,903

Lambda	K	H
1.37	0.711	1.31

Gapped

Lambda	K	H
1.37	0.711	1.31

Matrix: blastn matrix:1 -3

Gap Penalties: Existence: 5, Extension: 2

Number of Hits to DB: 6901

Number of Sequences: 1205903

Number of extensions: 6901

Number of successful extensions: 1479

Number of sequences better than 10.0: 16

length of query: 16

length of database: 5,297,768,116

effective HSP length: 15

effective length of query: 1  
effective length of database: 5,279,679,571  
effective search space: 5279679571  
effective search space used: 5279679571  
T: 0  
A: 30  
X1: 6 (11.9 bits)  
X2: 15 (29.7 bits)  
S1: 12 (24.3 bits)  
S2: 15 (30.2 bits)

## Top of Form

1: NM\_000660. Homo sapiensRelated Sequences, OMIM, Protein, PubMed,  
tran...[gi:10863872] Taxonomy, UniSTS, LinkOut

LOCUS NM\_000660 2745 bp mRNA linear PRI 13-F  
EB-2002

DEFINITION Homo sapiens transforming growth factor, beta 1 (Camurati-En  
gelmann

disease) (TGFB1), mRNA.

ACCESSION NM\_000660

VERSION NM\_000660.1 GI:10863872

KEYWORDS .

SOURCE human.

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleos  
tomi;

Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 2745)

- AUTHORS Derynck, R., Jarrett, J.A., Chen, E.Y., Eaton, D.H., Bell, J.R.,  
Assoian, R.K., Roberts, A.B., Sporn, M.B. and Goeddel, D.V.
- TITLE Human transforming growth factor-beta complementary DNA sequ  
ence  
and expression in normal and transformed cells
- JOURNAL Nature 316 (6030), 701-705 (1985)
- MEDLINE 85296301
- REFERENCE 2 (bases 1 to 2745)
- AUTHORS Sporn, M.B., Roberts, A.B., Wakefield, L.M. and Assoian, R.K.
- TITLE Transforming growth factor-beta: biological function and che  
mical  
structure
- JOURNAL Science 233 (4763), 532-534 (1986)
- MEDLINE 86261803
- PUBMED 3487831
- REFERENCE 3 (bases 1 to 2745)
- AUTHORS Chang, N.S., Mattison, J., Cao, H., Pratt, N., Zhao, Y. and Lee, C
- TITLE Cloning and characterization of a novel transforming growth  
factor-beta1-induced TIAF1 protein that inhibits tumor necro  
sis  
factor cytotoxicity
- JOURNAL Biochem. Biophys. Res. Commun. 253 (3), 743-749 (1998)
- MEDLINE 99119079
- PUBMED 9918798
- REFERENCE 4 (bases 1 to 2745)
- AUTHORS Ghadami, M., Makita, Y., Yoshida, K., Nishimura, G., Fukushima, Y  
.,  
Wakui, K., Ikegawa, S., Yamada, K., Kondo, S., Niikawa, N. and To

mita, H.

TITLE Genetic mapping of the Camurati-Engelmann disease locus to  
chromosome 19q13.1-q13.3  
JOURNAL Am. J. Hum. Genet. 66 (1), 143-147 (2000)  
MEDLINE 20100617  
PUBMED 10631145  
REFERENCE 5 (bases 1 to 2745)  
AUTHORS Vaughn, S.P., Broussard, S., Hall, C.R., Scott, A., Blanton, S.H.

Milunsky, J.M. and Hecht, J.T.

TITLE Confirmation of the mapping of the Camurati-Engelmann locus  
to

19q13. 2 and refinement to a 3.2-cM region  
JOURNAL Genomics 66 (1), 119-121 (2000)  
MEDLINE 20304762  
PUBMED 10843814  
REFERENCE 6 (bases 1 to 2745)  
AUTHORS Lim, J.M., Kim, J.A., Lee, J.H. and Joo, C.K.  
TITLE Downregulated expression of integrin alpha6 by transforming  
growth

factor-beta(1) on lens epithelial cells in vitro  
JOURNAL Biochem. Biophys. Res. Commun. 284 (1), 33-41 (2001)  
MEDLINE 21268957  
PUBMED 11374867  
COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to  
final  
NCBI review. The reference sequence was derived from X02812.

1.

FEATURES Location/Qualifiers

source 1..2745  
/organism="Homo sapiens"  
/db\_xref="taxon:9606"  
/chromosome="19"  
/map="19q13.1"

gene 1..2745  
/gene="TGFB1"  
/note="TGFB; DPD1; CED"  
/db\_xref="LocusID:7040"  
/db\_xref="MIM:190180"

misc\_feature 37..113  
/note="pot. hairpin loops-forming region"

variation 72  
/allele="-"  
/allele="C"  
/db\_xref="dbSNP:1800999"

variation 79  
/allele="-"  
/allele="C"  
/db\_xref="dbSNP:1799753"

CDS 842..2017  
/gene="TGFB1"  
/note="transforming growth factor, beta 1; diaphyse  
al  
dysplasia 1, progressive (Camurati-Engelmann diseas  
e)"  
/codon\_start=1  
/db\_xref="LocusID:7040"  
/db\_xref="MIM:190180"

/product="transforming growth factor, beta 1  
(Camurati-Engelmann disease)"  
/protein\_id="NP\_000651.1"  
/db\_xref="GI:10863873"  
/translation="MPPSGLRLLPLLPLWLLVLTGPPPAAGLSTCKTID

MELVKRK

RIEAIRGQILSKRLASPPSQGEVPPGPLPEAVLALYNSTRDRVAGESAEPEPEPEAD  
YYAKEVTRVLMVETHNEIYDKFKQSTHSIYMFFNTSELREAVPEPVLLSRAELRLRR  
LKLKVEQHVELYQKYSNNSWRYLSNRLAPSDSPEWLSFDVTGVVRQWLSRGGEIEGF  
RLSAHCSCDSRDNTLQVDINGFTTGRRGDLATIHGMNRPFLLLMATPLERAQHLQSSR  
HRRALDTNYCFSSTEKNCCVRQLYIDFRKDLGWKWIHEPKGYHANFCLGPCPYIWSLD  
TQYSKVLALYNQHNPGASAAPCCVPQALEPLIVYYYVGRKPKVEQLSNMIVRSCKCS"

misc\_feature 863..910

/note="pot. core sequence of signal peptide (aa -27

2 to

-257)"

variation 870

/allele="C"

/allele="T"

/db\_xref="dbSNP:1982073"

variation 915

/allele="C"

/allele="G"

/db\_xref="dbSNP:1800471"

misc\_feature 938..1600

/note="TGFb\_propeptide; Region: TGF-beta propeptide

"

misc\_feature 953

/note="pot. altern. translation start site"



misc\_feature 1035..1043  
/note="put. glycosylation site"

misc\_feature 1247..1255  
/note="put. glycosylation site"

misc\_feature 1370..1378  
/note="put. glycosylation site"

variation 1632  
/allele="C"  
/allele="T"  
/db\_xref="dbSNP:1800472"

mat\_peptide 1679..2014  
/product="mature TGF-beta (aa 1-112)"

misc\_feature 1715..2014  
/note="TGF-beta; Region: Transforming growth factor  
beta  
like domain"

misc\_feature 1721..2014  
/note="TGFB; Region: Transforming growth factor-beta  
a  
(TGF-beta) family"

misc\_feature 2018..2096  
/note="GC-rich region"

promoter 2097..2103  
/note="TATA-box-like region"

misc\_feature 2517..2522  
/note="put. polyadenylation signal"

polyA\_site 2539  
/note="polyadenylation site"

BASE COUNT 527 a 938 c 801 g 479 t

## ORIGIN

1 acctccctcc gcggagcagc cagacagcga gggccccggc cgggggcagg ggggacg  
ccc  
61 cgtccggggc acccccccg gctctgagcc gcccgcgggg ccggcctcgg cccggag  
cgg  
121 aggaaggagt cgccgaggag cagcctgagg cccagagtc tgagacgagc cgccgcc  
gcc  
181 cccgccactg cggggaggag ggggaggagg agcgggagga gggacgagct ggtcggg  
aga  
241 agaggaaaaa aacttttgag acttttccgt tgccgctggg agccggaggc gcgggga  
cct  
301 cttggcgca cgctgccccg cgaggaggca ggacttgggg acccagacc gcctccc  
ttt  
361 gccgccgggg acgttgctc cctccctgcc ccctacacgg cgtccctcag gcgcccc  
cat  
421 tccggaccag ccctcgggag tcgccgaccc ggcctcccgc aaagactttt cccaga  
cct  
481 cgggcgcacc ccctgcacgc cgccttcac cccggcctgt ctctgagcc cccgcgc  
atc  
541 ctagaccctt tctctccag gagacggatc tctctccgac ctgccacaga tccccta  
ttc  
601 aagaccaccc accttctggt accagatgc gcccatctag gttatttccg tgggata  
ctg  
661 agacaccccc ggtccaagcc tcccctccac cactgcgcc ttctccctga ggagcct  
cag  
721 ctttccctcg aggccctcct accttttgcc gggagacccc cagcccctgc aggggcg  
ggg  
781 cctccccacc acaccagccc tgttcgcgt ctcggcagt cggggggcg ccgcctc  
ccc

841 catgccgccc tccgggctgc ggctgctgcc gctgctgcta ccgctgctgt ggctact  
ggt  
901 gctgacgcct ggcccgcgg ccgcgggact atccacctgc aagactatcg acatgga  
gct  
961 ggtgaagcgg aagcgcacgc aggccatccg cggccagatc ctgtccaagc tgcggct  
cgc  
1021 cagccccccg agccagggggg aggtgccgcc cggcccgtg cccgaggccg tgctcgc  
cct  
1081 gtacaacagc acccgcgacc ggggtggccgg ggagagtgc gaaccggagc ccgagcc  
tga  
1141 ggccgactac tacgccaagg aggtcaccgc cgtgctaata gtggaaacct acaacga  
aat  
1201 ctatgacaag ttcaagcaga gtacacacag catatatatg ttcttcaaca catcaga  
gct  
1261 ccgagaagcg gtacctgaac ccgtgttgc ctcccgggca gagctgcgtc tgctgag  
gag  
1321 gctcaagtta aaagtggagc agcacgtgga gctgtaccag aaatacagca acaattc  
ctg  
1381 gcgatacctc agcaaccggc tgctggcacc cagcgactcg ccagagtggg tatcttt  
tga  
1441 tgtcaccgga gttgtgcggc agtggttgag ccgtggaggg gaaattgagg gctttcg  
cct  
1501 tagcggccac tgctcctgtg acagcagga taacacactg caagtggaca tcaacgg  
gtt  
1561 cactaccggc cgccgaggtg acctggccac cattcatggc atgaaccggc ctttcct  
gct  
1621 tctcatggcc accccgctgg agagggccca gcatctgcaa agctcccggc accgccg  
agc  
1681 cctggacacc aactattgct tcagctccac ggagaagaac tgctgcgtgc ggcagct

gta

1741 cattgacttc cgcaaggacc tcggctggaa gtggatccac gagcccaagg gctacca

tgc

1801 caacttctgc ctcgggccct gcccctacat ttggagcctg gacacgcagt acagcaa

ggt

1861 cctggccctg tacaaccagc ataaccggg cgcctcggcg gcgccgtgct gcgtgcc

gca

1921 ggcgctggag ccgctgcccc tcgtgtacta cgtgggcccgc aagcccaagg tggagca

gct

1981 gtccaacatg atcgtgcgct cctgcaagtg cagctgaggt cccgccccgc cccgccc

cgc

2041 cccggcaggc ccggccccac cccgccccgc ccccgtgcc ttgcccatgg gggctgt

att

2101 taaggacacc gtgccccaaag cccacctggg gcccattaa agatggagag aggactg

cgg

2161 atctctgtgt cattgggcgc ctgcctgggg tctccatccc tgacgttccc ccactcc

cac

2221 tccctctctc tccctctctg cctcctcctg cctgtctgca ctattccttt gcccggc

atc

2281 aaggcacagg ggaccagtgg ggaacactac tgtagttaga tctatttatt gagcacc

ttg

2341 ggcactgttg aagtgcctta cattaatgaa ctcatcagt caccatagca acactct

gag

2401 atggcaggga ctctgataac acccatttta aaggttgagg aaacaagccc agagagg

tta

2461 agggaggagt tcctgccac caggaacctg ctttagtggg ggatagtga gaagaca

ata

2521 aaagatagta gttcaggcca ggcggggtgc tcacgcctgt aatcctagca cttttgg

gag

2581 gcagagatgg gaggatactt gaatccaggc atttgagacc agcctgggta acatagt  
gag

2641 accctatctc tacaaaacac ttttaaaaaa tgtacacctg tgggtcccagc tactctg  
gag

2701 gctaaggtgg gaggatcact tgatcctggg aggtcaaggc tgcag  
//

Bottom of Form

Revised: October 24, 2001.

Query= (16 letters)

Database: GenBank Human EST entries

4,280,058 sequences; 2,114,234,064 total letters

Score E

Sequences producing significant alignments: (bits)

Value

gi|19365764|gb|BM915385.1|BM915385 AGENCOURT\_6701642 NIH\_MG... 32

0.41

gi|19353768|gb|BM903897.1|BM903897 AGENCOURT\_6696012 NIH\_MG... 32

0.41

gi|18807810|gb|BM562052.1|BM562052 AGENCOURT\_6562015 NIH\_MG... 32

0.41

gi|18791603|gb|BM553137.1|BM553137 AGENCOURT\_6572574 NIH\_MG... 32

0.41

gi|16171065|gb|BI908151.1|BI908151 603067456F1 NIH\_MGC\_118 ... 32

0.41

gi|15759271|gb|BI767693.1|BI767693 603060648F1 NIH\_MGC\_122 ... 32

0.41

gi|15343643|gb|BI518851.1|BI518851 603061760F1 NIH\_MGC\_118 ... 32

0.41

gi 14309343 gb BG899094.1 BG899094	HOA21-1-G9 HOA (Human Os...	32
0.41		
gi 13662542 gb BG611171.1 BG611171	602612144F1 NIH_MGC_60 H...	32
0.41		
gi 12609210 gb BG115704.1 BG115704	602317174F1 NIH_MGC_88 H...	32
0.41		
gi 12101282 gb BF796228.1 BF796228	602258513F1 NIH_MGC_85 H...	32
0.41		
gi 11152079 gb BF238160.1 BF238160	601811886F1 NIH_MGC_48 H...	32
0.41		
gi 11100313 gb BF206727.1 BF206727	601871105F1 NIH_MGC_19 H...	32
0.41		
gi 11100272 gb BF206686.1 BF206686	601871051F1 NIH_MGC_19 H...	32
0.41		
gi 16775383 gb BM046103.1 BM046103	603625849F1 NIH_MGC_40 H...	30
1.6		
gi 19739174 gb BQ014273.1 BQ014273	UI-H-ED1-axs-h-21-0-UI.s...	28
6.4		
gi 19378603 gb BM928224.1 BM928224	AGENCOURT_6699855 NIH_MG...	28
6.4		
gi 19367808 gb BM917429.1 BM917429	AGENCOURT_6606724 NIH_MG...	28
6.4		
gi 19364214 gb BM913835.1 BM913835	AGENCOURT_6612786 NIH_MG...	28
6.4		
gi 19361343 gb BM910964.1 BM910964	AGENCOURT_6615957 NIH_MG...	28
6.4		
gi 18505954 gb BM456914.1 BM456914	AGENCOURT_6404253 NIH_MG...	28
6.4		
gi 18499709 gb BM450669.1 BM450669	AGENCOURT_6394717 NIH_MG...	28

6.4

gi|16000196|gb|BI859449.1|BI859449 603388188F1 NIH\_MGC\_87 H... 28

6.4

gi|15928460|gb|BI818193.1|BI818193 603032663F1 NIH\_MGC\_115 ... 28

6.4

gi|15431547|gb|BI544235.1|BI544235 603241605F1 NIH\_MGC\_95 H... 28

6.4

gi|15345229|gb|BI520437.1|BI520437 603071622F1 NIH\_MGC\_119 ... 28

6.4

gi|14440373|gb|BI033747.1|BI033747 PM3-NN0223-220201-014-h0... 28

6.4

gi|14426676|gb|BI020046.1|BI020046 CM3-MT0291-110101-622-f0... 28

6.4

gi|14081325|gb|BG770672.1|BG770672 602734012F1 NIH\_MGC\_49 H... 28

6.4

gi|13546630|gb|BG547965.1|BG547965 602576071F1 NIH\_MGC\_77 H... 28

6.4

gi|13030375|gb|BG281450.1|BG281450 602401966F1 NIH\_MGC\_20 H... 28

6.4

gi|12951460|emb|AL582959.1|AL582959 AL582959 LTI\_NFL010\_BC2... 28

6.4

gi|12764352|gb|BG254536.1|BG254536 602368464F1 NIH\_MGC\_91 H... 28

6.4

gi|12378592|gb|BF961317.1|BF961317 PM3-NN0223-111200-004-d0... 28

6.4

gi|12374538|gb|BF957263.1|BF957263 PM3-NN0223-241100-002-b0... 28

6.4

gi|12323114|gb|BF926150.1|BF926150 CM2-NT0193-301100-562-a1... 28

6.4

gi 12259862 gb BF869732.1 BF869732	IL3-ET0114-251000-316-A1...	28
6.4		
gi 12129894 gb BF800905.1 BF800905	PM1-CI0110-201000-003-f0...	28
6.4		
gi 12071436 gb BF744760.1 BF744760	QV2-BT0635-311000-440-c1...	28
6.4		
gi 11770407 gb BE965733.2 BE965733	601659792R1 NIH_MGC_70 H...	28
6.4		
gi 11766539 gb BE963121.2 BE963121	601656923R1 NIH_MGC_67 H...	28
6.4		
gi 10348536 gb BE890328.1 BE890328	601431783F1 NIH_MGC_72 H...	28
6.4		
gi 10142985 gb BE728993.1 BE728993	601562251F1 NIH_MGC_20 H...	28
6.4		
gi 10095527 gb BE707262.1 BE707262	PM1-HT0452-060700-008-e0...	28
6.4		
gi 9772196 gb BE543551.1 BE543551	601070523F1 NIH_MGC_12 Ho...	28
6.4		
gi 9768571 gb BE539926.1 BE539926	601060667F2 NIH_MGC_10 Ho...	28
6.4		
gi 9342607 gb BE397242.1 BE397242	601290754F1 NIH_MGC_8 Hom...	28
6.4		
gi 9332870 gb BE387505.1 BE387505	601274247F1 NIH_MGC_20 Ho...	28
6.4		
gi 8140649 gb AW950985.1 AW950985	EST363055 MAGE resequence...	28
6.4		
gi 8139665 gb AW950129.1 AW950129	EST362094 MAGE resequence...	28
6.4		
gi 6879658 gb AW375004.1 AW375004	MR0-CT0068-280999-002-f07...	28



6.4

gi|5435227|emb|AL079651.1|AL079651 DKFZp434N0629\_r1 434 (sy... 28

6.4

gi|5406349|emb|AL036861.1|AL036861 DKFZp56401963\_r1 564 (sy... 28

6.4

gi|2566893|gb|AA641675.1|AA641675 nr62g01.s1 NCI\_CGAP\_Lym3 ... 28

6.4

gi|2080087|gb|AA418268.1|AA418268 zv96d09.s1 Soares\_NhHMPu... 28

6.4

gi|2056455|gb|AA402650.1|AA402650 zu49g06.r1 Soares ovary t... 28

6.4

gi|1516398|gb|AA040102.1|AA040102 zk46e02.r1 Soares\_pregnan... 28

6.4

## Alignments

>gi|19365764|gb|BM915385.1|BM915385 AGENCOURT\_6701642 NIH\_MGC\_41 Homo s  
apiens cDNA clone

IMAGE:5481560 5'.

Length = 1086

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgaggag 16

|||||

Sbjct: 23 acctccctccgaggag 38

>gi|19353768|gb|BM903897.1|BM903897 AGENCOURT\_6696012 NIH\_MGC\_67 Homo s  
apiens cDNA clone IMAGE:5492392

5'.

Length = 1497

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 445 acctccctccgcggag 430

>gi|18807810|gb|BM562052.1|BM562052 AGENCOURT\_6562015 NIH\_MGC\_118 Homo  
sapiens cDNA clone

IMAGE:5745414 5'.

Length = 1175

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 20 acctccctccgcggag 35

>gi|18791603|gb|BM553137.1|BM553137 AGENCOURT\_6572574 NIH\_MGC\_41 Homo s  
apiens cDNA clone

IMAGE:5467063 5'.

Length = 1100

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

||||||||||||||

Sbjct: 26 acctccctccgcggag 41

>gi|16171065|gb|BI908151.1|BI908151 603067456F1 NIH\_MGC\_118 Homo sapien  
s cDNA clone IMAGE:5216508 5'.

Length = 706

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

||||||||||||||

Sbjct: 25 acctccctccgcggag 40

>gi|15759271|gb|BI767693.1|BI767693 603060648F1 NIH\_MGC\_122 Homo sapien  
s cDNA clone IMAGE:5209978 5'.

Length = 862

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 705 acctccctccgcggag 720

>gi|15343643|gb|BI518851.1|BI518851 603061760F1 NIH\_MGC\_118 Homo sapien  
s cDNA clone IMAGE:5210943 5'.

Length = 943

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 25 acctccctccgcggag 40

>gi|14309343|gb|BG899094.1|BG899094 HOA21-1-G9 HOA (Human Osteoarthritis  
c Cartilage) Homo sapiens

cDNA.

Length = 364

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 83 acctccctccgcggag 98

>gi|13662542|gb|BG611171.1|BG611171 602612144F1 NIH\_MGC\_60 Homo sapiens

cDNA clone IMAGE:4737466 5'.

Length = 897

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 809 acctccctccgcggag 794

>gi|12609210|gb|BG115704.1|BG115704 602317174F1 NIH\_MGC\_88 Homo sapiens  
cDNA clone IMAGE:4417482 5'.

Length = 838

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 51 acctccctccgcggag 66

>gi|12101282|gb|BF796228.1|BF796228 602258513F1 NIH\_MGC\_85 Homo sapiens  
cDNA clone IMAGE:4341962 5'.

Length = 1081

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 7 acctccctccgcggag 22

>gi|11152079|gb|BF238160.1|BF238160 601811886F1 NIH\_MGC\_48 Homo sapiens

cDNA clone IMAGE:4054821 5'.

Length = 811

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 11 acctccctccgcggag 26

>gi|11100313|gb|BF206727.1|BF206727 601871105F1 NIH\_MGC\_19 Homo sapiens

cDNA clone IMAGE:4101600 5'.

Length = 888

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

|||||

Sbjct: 32 acctccctccgcggag 47

>gi|11100272|gb|BF206686.1|BF206686 601871051F1 NIH\_MGC\_19 Homo sapiens  
cDNA clone IMAGE:4101517 5'.

Length = 917

Score = 32.2 bits (16), Expect = 0.41

Identities = 16/16 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcggag 16

||||||||||||||

Sbjct: 33 acctccctccgcggag 48

>gi|16775383|gb|BM046103.1|BM046103 603625849F1 NIH\_MGC\_40 Homo sapiens  
cDNA clone IMAGE:5452309 5'.

Length = 869

Score = 30.2 bits (15), Expect = 1.6

Identities = 15/15 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcggag 16

||||||||||||||

Sbjct: 692 cctccctccgcggag 706

>gi|19739174|gb|BQ014273.1|BQ014273 UI-H-ED1-axs-h-21-0-UI.s1 NCI\_CGAP\_  
ED1 Homo sapiens cDNA clone

IMAGE:5833028 3'.

Length = 772

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 495 cctccctccgcgga 482

>gi|19378603|gb|BM928224.1|BM928224 AGENCOURT\_6699855 NIH\_MGC\_121 Homo  
sapiens cDNA clone IMAGE:5770072 5'.

Length = 1140

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 1009 cctccctccgcgga 1022

>gi|19367808|gb|BM917429.1|BM917429 AGENCOURT\_6606724 NIH\_MGC\_106 Homo  
sapiens cDNA clone IMAGE:5483947 5'.

Length = 1073

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus



Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 916 acctccctccgcgg 929

>gi|19364214|gb|BM913835.1|BM913835 AGENCOURT\_6612786 NIH\_MGC\_98 Homo s  
apiens cDNA clone IMAGE:5477539 5'.

Length = 1104

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 842 cctccctccgcgga 829

>gi|19361343|gb|BM910964.1|BM910964 AGENCOURT\_6615957 NIH\_MGC\_98 Homo s  
apiens cDNA clone IMAGE:5454547 5'.

Length = 1128

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 3 ctccctccgcggag 16

||||||||||||

Sbjct: 883 ctccctccgcggag 870

>gi|18505954|gb|BM456914.1|BM456914 AGENCOURT\_6404253 NIH\_MGC\_92 Homo sapiens cDNA clone

IMAGE:5583862 5'.

Length = 1813

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 29 cctccctccgcgga 42

>gi|18499709|gb|BM450669.1|BM450669 AGENCOURT\_6394717 NIH\_MGC\_67 Homo sapiens cDNA clone IMAGE:5494366 5'.

Length = 1430

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 1150 acctccctccgcgg 1163

>gi|16000196|gb|BI859449.1|BI859449 603388188F1 NIH\_MGC\_87 Homo sapiens cDNA clone IMAGE:5396997 5'.

Length = 852

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 100 acctccctccgcgg 113

>gi|15928460|gb|BI818193.1|BI818193 603032663F1 NIH\_MGC\_115 Homo sapien  
s cDNA clone IMAGE:5173838 5'.

Length = 683

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 96 cctccctccgcgga 109

>gi|15431547|gb|BI544235.1|BI544235 603241605F1 NIH\_MGC\_95 Homo sapiens  
cDNA clone IMAGE:5284296 5'.

Length = 676

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 3 ctccctccgcggag 16

||||||||||||

Sbjct: 39 ctccctccgcggag 26

>gi|15345229|gb|BI520437.1|BI520437 603071622F1 NIH\_MGC\_119 Homo sapien  
s cDNA clone IMAGE:5163773 5'.

Length = 727

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 505 acctccctccgcgg 492

>gi|14440373|gb|BI033747.1|BI033747 PM3-NN0223-220201-014-h04 NN0223 Ho  
mo sapiens cDNA.

Length = 284

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 97 acctccctccgcgg 84

>gi|14426676|gb|BI020046.1|BI020046 CM3-MT0291-110101-622-f04 MT0291 Ho  
mo sapiens cDNA.

Length = 436

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 365 acctccctccgcgg 352

>gi|14081325|gb|BG770672.1|BG770672 602734012F1 NIH\_MGC\_49 Homo sapiens  
cDNA clone IMAGE:4859546 5'.

Length = 949

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 63 acctccctccgcgg 76

>gi|13546630|gb|BG547965.1|BG547965 602576071F1 NIH\_MGC\_77 Homo sapiens  
cDNA clone IMAGE:4704209 5'.

Length = 918

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 248 acctccctccgcgg 261

>gi|13030375|gb|BG281450.1|BG281450 602401966F1 NIH\_MGC\_20 Homo sapiens  
cDNA clone IMAGE:4544201 5'.

Length = 782

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 417 acctccctccgcgg 430

>gi|12951460|emb|AL582959.1|AL582959 AL582959 LTI\_NFL010\_BC2 Homo sapie  
ns cDNA clone CS0DL008YA12 3 prime.

Length = 822

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 533 cctccctccgcgga 520

>gi|12764352|gb|BG254536.1|BG254536 602368464F1 NIH\_MGC\_91 Homo sapien  
s cDNA clone IMAGE:4476902 5'.

Length = 1031

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 849 cctccctccgcgga 862

>gi|12378592|gb|BF961317.1|BF961317 PM3-NN0223-111200-004-d03 NN0223 H  
omo sapiens cDNA.

Length = 277

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 89 acctccctccgcgg 76

>gi|12374538|gb|BF957263.1|BF957263 PM3-NN0223-241100-002-b08 NN0223 H  
omo sapiens cDNA.

Length = 168

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 117 acctccctccgcgg 104

>gi|12323114|gb|BF926150.1|BF926150 CM2-NT0193-301100-562-a12 NT0193 Homo sapiens cDNA.

Length = 417

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 268 cctccctccgcgga 255

>gi|12259862|gb|BF869732.1|BF869732 IL3-ET0114-251000-316-A11 ET0114 Homo sapiens cDNA.

Length = 278

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||



Sbjct: 73 acctccctccgcgg 60

>gi|12129894|gb|BF800905.1|BF800905 PM1-CI0110-201000-003-f08 CI0110 Homo sapiens cDNA.

Length = 283

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

|||||

Sbjct: 211 acctccctccgcgg 224

>gi|12071436|gb|BF744760.1|BF744760 QV2-BT0635-311000-440-c11 BT0635 Homo sapiens cDNA.

Length = 534

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

|||||

Sbjct: 319 cctccctccgcgga 332

>gi|11770407|gb|BE965733.2|BE965733 601659792R1 NIH\_MGC\_70 Homo sapiens cDNA clone IMAGE:3896134 3'.

Length = 1336

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 292 acctccctccgcgg 305

>gi|11766539|gb|BE963121.2|BE963121 601656923R1 NIH\_MGC\_67 Homo sapiens

cDNA clone IMAGE:3865924 3'.

Length = 1442

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 3 ctccctccgcggag 16

||||||||||||

Sbjct: 403 ctccctccgcggag 390

>gi|10348536|gb|BE890328.1|BE890328 601431783F1 NIH\_MGC\_72 Homo sapiens

cDNA clone IMAGE:3916820 5'.

Length = 794

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 115 acctccctccgcgg 128

>gi|10142985|gb|BE728993.1|BE728993 601562251F1 NIH\_MGC\_20 Homo sapiens  
cDNA clone IMAGE:3831924 5'.

Length = 840

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 397 acctccctccgcgg 410

>gi|10095527|gb|BE707262.1|BE707262 PM1-HT0452-060700-008-e08 HT0452 Ho  
mo sapiens cDNA.

Length = 592

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Minus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 343 acctccctccgcgg 330

>gi|9772196|gb|BE543551.1|BE543551 601070523F1 NIH\_MGC\_12 Homo sapiens

cDNA clone IMAGE:3456940 5'.

Length = 1035

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

|||||

Sbjct: 332 acctccctccgcgg 345

>gi|9768571|gb|BE539926.1|BE539926 601060667F2 NIH\_MGC\_10 Homo sapiens

cDNA clone IMAGE:3447161 5'.

Length = 902

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

|||||

Sbjct: 411 acctccctccgcgg 424

>gi|9342607|gb|BE397242.1|BE397242 601290754F1 NIH\_MGC\_8 Homo sapiens c

DNA clone IMAGE:3621253 5'.

Length = 524

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

||||||||||||

Sbjct: 228 cctccctccgcgga 241

>gi|9332870|gb|BE387505.1|BE387505 601274247F1 NIH\_MGC\_20 Homo sapiens  
cDNA clone IMAGE:3615538 5'.

Length = 637

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 1 acctccctccgcgg 14

||||||||||||

Sbjct: 422 acctccctccgcgg 435

>gi|8140649|gb|AW950985.1|AW950985 EST363055 MAGE resequences, MAGA Hom  
o sapiens cDNA.

Length = 638

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

|||||

Sbjct: 273 cctccctccgcgga 286

>gi|8139665|gb|AW950129.1|AW950129 EST362094 MAGE resequences, MAGA Homo sapiens cDNA.

Length = 611

Score = 28.2 bits (14), Expect = 6.4

Identities = 14/14 (100%)

Strand = Plus / Plus

Query: 2 cctccctccgcgga 15

|||||

Sbjct: 273 cctccctccgcgga 286

Database: GenBank Human EST entries

Posted date: Mar 29, 2002 2:35 AM

Number of letters in database: 2,114,234,064

Number of sequences in database: 4,280,058

Lambda	K	H
1.37	0.711	1.31

Gapped

Lambda	K	H
1.37	0.711	1.31

Matrix: blastn matrix:1 -3

Gap Penalties: Existence: 5, Extension: 2

Number of Hits to DB: 5013

Number of Sequences: 4280058

Number of extensions: 5013  
Number of successful extensions: 5013  
Number of sequences better than 10.0: 61  
length of query: 16  
length of database: 2,114,234,064  
effective HSP length: 15  
effective length of query: 1  
effective length of database: 2,050,033,194  
effective search space: 2050033194  
effective search space used: 2050033194  
T: 0  
A: 30  
X1: 6 (11.9 bits)  
X2: 15 (29.7 bits)  
S1: 12 (24.3 bits)  
S2: 14 (28.2 bits)  
Top of Form

1: BM915385.AGENCOURT\_6701642...[gi:19365764]

Taxono

my,

LinkO

ut

#### IDENTIFIERS

dbEST Id: 11598757  
EST name: AGENCOURT\_6701642  
GenBank Acc: BM915385  
GenBank gi: 19365764  
CLONE INFO  
Clone Id: IMAGE:5481560 (5')  
Plate: LLCM2006 Row: d Column: 09

DNA type: cDNA

PRIMERS

PolyA Tail: Unknown

SEQUENCE

CGCCCTGGGCCATCTCCCTCCCACCTCCCTCCGCGGAGCAGCCAGACAGCGAGGGC  
 CCGG  
 GCCGGGGGCAGGGGGGACGCCCCGTCCGGGGCACCCCCCGGCTCTGAGCCGCCCG  
 CGGG  
 GCCGGCCTCGGCCCGGAGCGGAGGAAGGAGTCGCCGAGGAGCAGCCTGAGGCCCA  
 GAGT  
 CTGAGACGAGCCGCCGCCGCCGCCACTGCGGGGAGGAGGGGAGGAGGAGCGG  
 GAGG  
 AGGGACGAGCTGGTCGGGAGAAGAGGAAAAAACTTTTGAGACTTTTCCGTTGCCG  
 CTGG  
 GAGCCGGAGGCGCGGGGACCTCTTGGCGGACGCTGCCCCGCGAGGAGGCAGGACT  
 TGGG  
 GACCCCAGACCGCCTCCCTTTGCCGCCGGGGACGCTTGCTCCCTCCCTGCCCCCTA  
 CACG  
 GCGTCCCTCAGGCGCCCCCATTCGGACCAGCCCTCGGGAGTCGCCGACCCGGCCT  
 CTCG  
 CAAAGACTTTTCACCATACTCGGGCGCACCTCTGCACGGGCCTTCATCACCGG  
 CCTG  
 TCTACTGAGCCCCCGCGGATGCCTAGACCCTTTCTCCTCCGGGAGACGGATCCCTC  
 TCCG  
 ACCTGCCGCAAATTCCTATTCTGGAACACCCCGCTTCCTGGGACCCTAATCCCC  
 GCCT  
 TTCGACGCTCCTTGCGCTGGGGAAGTGAAGAGCCCCCGGGTTCGTAACCTTTTCCT  
 TCCC  
 CGTTTTGAAAAACATCCCCCGTTAATAAACCTTGACTATTTTCGCTTTGGGCCCCC



CCCT

TACGGTTTTTGGCGGGCACTAAACAAACATCGAGTCTCAAGGCGGCGGATGCCACT

CAAG

CCTGAATACTTTTGC GCGTTAGGGGCGGTCTTTTACGCGAGTAGAGTCGGGCCTTG

ACCG

GACCCTATTCATTGGTTTCCCGTGACGTGTGCGGGCGTAACGAGATATTAACCTCT

CCCG

ACACATTGTCATAAAACACCACTTTCGACACGCCCTACTCCTGTTAATAGTCGCCC

CCTC

CCCGCGTGTA AAAATTTCCCGCGCCAATGCCCTCCATTATTCCGCTCCATGAAAAAG

GGGG

TCGGCN

Quality: High quality sequence stops at base: 467

Entry Created: Mar 11 2002

Last Updated: Mar 12 2002

## COMMENTS

Tissue Procurement: DCTD/DTP

cDNA Library Preparation: Rubin Laboratory

cDNA Library Arrayed by: The I.M.A.G.E. Consortium (LLNL

)

DNA Sequencing by: Agencourt Bioscience Corporation

Clone distribution: MGC clone distribution information c

an

be found through the I.M.A.G.E. Consortium/LLNL at:

<http://image.llnl.gov>

LIBRARY

Lib Name: NIH\_MGC\_41  
Organism: Homo sapiens  
Organ: skin  
Tissue type: amelanotic melanoma, cell line  
Lab host: DH10B (phage-resistant)  
Vector: pOTB7  
R. Site 1: XhoI  
R. Site 2: EcoRI  
Description: cDNA made by oligo-dT priming. Directionally cloned into  
EcoRI/XhoI sites using the following 5' adaptor: GGCACGA  
G(G  
) . Library constructed by Ling Hong in the laboratory of  
Gerald M. Rubin (University of California, Berkeley) usi  
ng  
ZAP-cDNA synthesis kit (Stratagene) and Superscript II R  
T  
(Life Technologies). Note: this is a NIH\_MGC Library.

## SUBMITTER

Name: Robert Strausberg, Ph.D.  
E-mail: cgapbs-r@mail.nih.gov

## CITATIONS

Title: National Institutes of Health, Mammalian Gene Collection  
(MGC)  
Authors: NIH-MGC <http://mgc.nci.nih.gov/>  
Year: 1999  
Status: Unpublished

Bottom of Form

Revised: October 24, 2001.

Check on Est in Genbank:

Query= (1086 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,  
GSS, or phase 0, 1 or 2 HTGS sequences)

1,205,903 sequences; 5,297,768,116 total letters

Score E

Sequences producing significant alignments: (bits)

Value

gi|10863872|ref|NM\_000660.1| Homo sapiens transforming grow... 587

e-165

gi|18590091|ref|XM\_085882.1| Homo sapiens similar to transf... 587

e-165

gi|11424057|ref|XM\_008912.1| Homo sapiens transforming grow... 587

e-165

gi|7684381|gb|AC011462.4|AC011462 Homo sapiens chromosome 1... 587

e-165

gi|37097|emb|X05839.1|HSTGFGB1 Human transforming growth fa... 587

e-165

gi|37092|emb|X02812.1|HSTGFB1 Human mRNA for transforming g... 587

e-165

gi|340526|gb|J04431.1|HUMTGFB1PR Homo sapiens transforming ... 587

e-165

gi|12654682|gb|BC001180.1|BC001180 Homo sapiens, Similar to... 291

8e-76

gi 12652748 gb BC000125.1 BC000125	Homo sapiens, Similar to...	291
8e-76		
gi 18490115 gb BC022242.1	Homo sapiens, clone MGC:22008 IM...	153
4e-34		
gi 755044 gb M23703.1 PIGTGFB1A	Sus scrofa transforming gro...	129
6e-27		
gi 7650477 gb AF249327.1 AF249327	Rattus norvegicus TGF-bet...	66
8e-08		
gi 4416081 gb AF105069.1 AF105069	Rattus norvegicus transfo...	66
8e-08		
gi 2394170 gb AF015683.1 AF015683	Rattus norvegicus transfo...	66
8e-08		
gi 6755774 ref NM_011577.1	Mus musculus transforming growt...	64
3e-07		
gi 1161133 gb L42456.1 MUSTGF1G01	Mus musculus TGF-1 gene, ...	64
3e-07		
gi 3688423 emb AJ009862.1 MMU009862	Mus musculus mRNA for t...	64
3e-07		
gi 201947 gb M57902.1 MUSTGFB1	Mouse transforming growth fa...	64
3e-07		
gi 18042365 gb AC097483.3	Homo sapiens BAC clone RP11-146N...	44
0.30		
gi 17481821 ref XM_008785.3	Homo sapiens one cut domain, f...	44
0.30		
gi 12737997 ref XM_007116.2	Homo sapiens Zic family member...	44
0.30		
gi 6005961 ref NM_007129.1	Homo sapiens Zic family member ...	44
0.30		
gi 11065969 gb AF193855.1 AF193855	Homo sapiens zinc finger...	44

0.30		
gi 4758847 ref NM_004852.1	Homo sapiens one cut domain, fa...	44
0.30		
gi 15787728 emb AL355338.33 AL355338	Human DNA sequence fro...	44
0.30		
gi 4028591 gb AF104902.1 AF104902	Homo sapiens ZIC2 protein...	44
0.30		
gi 1531593 gb U50523.1 HSU50523	Human BRCA2 region, mRNA se...	44
0.30		
gi 4468940 emb Y18198.1 HSAY18198	Homo sapiens mRNA for ONE...	44
0.30		
gi 19067958 gb AY049805.1	Alopias pelagicus 5.8S ribosomal...	42
1.2		
gi 18025465 gb AY037858.1	Cercopithicine herpesvirus 15 st...	42
1.2		
gi 12039248 gb AC020659.5 AC020659	Homo sapiens chromosome ...	42
1.2		
gi 19909461 gb AC098709.3	Mus musculus clone RP23-1K14, co...	40
4.6		
gi 19921137 ref NM_135651.1	Drosophila melanogaster (CG47...	40
4.6		
gi 18376846 gb AC092198.2	Homo sapiens chromosome X clone ...	40
4.6		
gi 18467841 ref XM_078995.1	CG4751 (CG4751), mRNA	40
4.6		
gi 18376869 gb AC091898.2	Homo sapiens chromosome 5 clone ...	40
4.6		
gi 18030132 gb AC026695.5	Homo sapiens chromosome 5 clone ...	40
4.6		

gi 15887302 gb AC020914.8  Homo sapiens chromosome 19 clone...	40
4.6	
gi 14578122 gb AC092241.1 AC092241 Drosophila melanogaster,...	40
4.6	
gi 15292266 gb AY051978.1  Drosophila melanogaster LD44770 ...	40
4.6	
gi 15055218 gb AC060226.39  Homo sapiens 12 BAC RP11-101P14...	40
4.6	
gi 14389338 gb AC084282.6 AC084282 Oryza sativa chromosome ...	40
4.6	
gi 13677167 gb AC015977.9 AC015977 Homo sapiens clone RP11-...	40
4.6	
gi 9910225 ref NM_020179.1  Homo sapiens FN5 protein (FN5),...	40
4.6	
gi 10440613 gb AC069145.5 AC069145 Oryza sativa chromosome ...	40
4.6	
gi 10728714 gb AE003631.2 AE003631 Drosophila melanogaster ...	40
4.6	
gi 9246422 gb AF197137.1 AF197137 Homo sapiens FN5 protein ...	40
4.6	
gi 4190938 gb AC000091.1 AC000091 Homo sapiens Chromosome 2...	40
4.6	
gi 17431932 emb AL646085.1 AL646085 Ralstonia solanacearum ...	40
4.6	
gi 15073719 emb AL591785.1 SME591785 Sinorhizobium meliloti...	40
4.6	
gi 3628578 gb AC005115.1 AC005115 Drosophila melanogaster D...	40
4.6	
gi 3150432 gb U50080.1 LSU50080 Lymnaea stagnalis serotonin...	40

4.6

gi|8052359|emb|AL356592.1|SC9H11 Streptomyces coelicolor co... 40

4.6

gi|6624640|emb|AL034344.24|HS118B18 Human DNA sequence from... 40

4.6

gi|15528721|dbj|AP003296.3| Oryza sativa (japonica cultivar... 40

4.6

gi|15289781|dbj|AP003141.2| Oryza sativa (japonica cultivar... 40

4.6

gi|6069643|dbj|AP000616.1| Oryza sativa (japonica cultivar-... 40

4.6

gi|960285|gb|L46862.1|RATLAMB2G Rattus norvegicus laminin B... 40

4.6

gi|198704|gb|J03749.1|MUSLAMB2B Mouse laminin B2 gene, exon... 40

4.6

gi|198702|gb|J02930.1|MUSLAMB2A Mouse laminin B2 chain mRNA... 40

4.6

gi|198694|gb|J03484.1|MUSLAM2B Mouse laminin B2 chain mRNA,... 40

4.6

## Alignments

>gi|10863872|ref|NM\_000660.1| Homo sapiens transforming growth factor,  
beta 1 (Camurati-Engelmann

disease) (TGFB1), mRNA

Length = 2745

Score = 587 bits (296), Expect = e-165

Identities = 356/377 (94%), Gaps = 1/377 (0%)

Strand = Plus / Plus

Query: 246 cgagctggtcgggagaagaggnnnnnnncttttgagacttttccgttgccgctgggagcc  
305

|||||

Sbjct: 225 cgagctggtcgggagaagaggaaaaaacttttgagacttttccgttgccgctgggagcc  
284

Query: 306 ggaggcgcggggacctcttggcgcgacgctgccccgcgaggaggcaggacttggggaccc  
365

|||||

Sbjct: 285 ggaggcgcggggacctcttggcgcgacgctgccccgcgaggaggcaggacttggggaccc  
344

Query: 366 cagaccgcctccctttgccgccggggacgcttgctccctccctgccccctacacggcgtc  
425

|||||

Sbjct: 345 cagaccgcctccctttgccgccggggacgcttgctccctccctgccccctacacggcgtc  
404

Query: 426 cctcaggcgccccattccggaccagccctcgggagtcgccgaccggcctctcgcaaag  
485

|||||

Sbjct: 405 cctcaggcgccccattccggaccagccctcgggagtcgccgaccggcctcccgcaaag  
464

Query: 486 acttttcaccatacctcgggcgcaccctctgcacgcggccttcacacggcctgtctac  
545

|||||

Sbjct: 465 acttttcccagacctcgggcgcacccctgcacgcgccttcacacggcctgtctcc  
524



Query: 546 tgagccccgcggatgcctagaccctttctcctccgggagacggatccctctccgacctg  
605

||||||| || ||||||||||||||| ||||||| |||||||

Sbjct: 525 tgagccccgcgcat-cctagaccctttctcctccaggagacggatctctctccgacctg  
583

Query: 606 ccgcaaattccctattc 622

|| || || |||||

Sbjct: 584 ccacagatcccctattc 600

【 0 0 8 3 】

#### Example 6: Statistical analysis of 5' end sequence tags

5' end sequence tags obtained from the same plurality of mRNAs in a sample or nucleic acid fragments within the same cDNA library can be analyzed by a standard software solution like NCBI BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) to identify non-redundant sequence tags as describe in Example 5. All such non-redundant sequence tags can then be individually counted and further analyzed for the contribution of each non-redundant tag to the total number of all tags obtained from the same sample. The contribution of an individual tag to the total number of all tags should allow for a quantification of the transcripts in a plurality of mRNAs in the sample or a cDNA library. The results obtained in such a way on individual samples can be further compared with similar data obtained from other samples to compare their expression patterns.

【 0 0 8 4 】

#### Example 7: Mapping of 5' end sequence tags to genomic sequence information

5' end specific sequence tags obtained as describe in this Example can be used to identify transcribed regions within genomes for which partial

l or entire sequences were obtained. Such a search can be performed using standard software solutions like NCBI BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) to align the 5' end specific sequence tags to genomic sequences. In the case of large genomes like those from human, rat or mouse it may be necessary to extend the initial sequence information obtained from concatamers for example by the approach describe in Example #. The use of extended sequences allows for a more precise identification of actively transcribed regions in the genome.

#### 【 0 0 8 5 】

##### Example 8: Identification of transcriptional start sites

5' end specific sequence tags, which could be mapped to genomic sequences, allow for the identification of regulatory sequences. In a gene the DNA upstream of the 5' end of transcribed regions usually encompasses most of the regulatory elements, which are used in the control of gene expression. These regulatory sequences can be further analyzed for their functionality by searches in databases, which hold information on binding sites for transcription factors. Publicly available databases on transcription factor binding sites and for promoter analysis include:

Transcription Regulatory Region Database (TRRD) (<http://www.mgs.bionet.nsc.ru/mgs/dbases/trrd4/>)

TRANSFAC (<http://transfac.gbf.de/TRANSFAC/>)

TFSEARCH (<http://www.cbrc.jp/research/db/TFSEARCH.html>)

PromoterInspector provide by Genomatix Software (<http://www.genomatix.de/>)

#### 【 0 0 8 6 】

##### Example 9: Cloning of full-length cDNAs using information derived from 5' end sequence tags

Sequence information derived from the concatamers can be used to synth

esis specific primers for the cloning of full-length cDNAs. In such an approach, the sequence derived from a given 5' end specific tag can be used to design a forward primer while the choice of the reverse primer would be dependent on the template DNA used in the amplification reaction. Amplification by the polymerase chain reaction (PCR) can be performed using a template derived from a plurality of RNA obtained from a biological sample and an oligo-dT primer. In the first step the oligo-dT primer and a reverse transcriptase are used to synthesis a cDNA pool. In the second step a forward primer derived from a 5' end specific tag and an oligo-dT primer are used to amplify a full-length cDNA from the cDNA pool. Similarly, a specific full-length cDNA can be amplified from an existing cDNA library using a forward primer derived from a 5' end tag and a vector nested reversed primer.

【0 0 8 7】

Example 10: Alternative approaches for the cloning of 5'-end tags from cDNA libraries.

A plurality of cDNAs can be amplified from an existing cDNA library having a recognition site for a class IIs endonuclease at the 5' end of the inserts. The PCR products derived from such a library would be further treated as described in the examples herein.

【0 0 8 8】

Example 11: Cloning of 5' ends by replacement of the Cap structure by an oligonucleotide having a class IIs recognition site

A cDNA/RNA hybrid encompassing the 5' end of an initial transcript can be obtained as described in Example 1. The Cap structure in such cDNA/RNA hybrids is then enzymatically removed by a hydrolyzing enzyme such as the T4 polynucleotide kinase or the tobacco acid pyrophosphatase. A single or double stranded oligonucleotide having a class IIs recognition site is then ligated by T4 RNA ligase to the RNA at the phosphate present

at the 5' end of the de-capped mRNA. The ligated oligonucleotide will function as a primer for the second strand synthesis following the procedure given in Example 1. By the use of a modified oligonucleotide in the ligation step the double stranded cDNA can be attached to a support and used for the cloning of concatamers as described herein.

【 0 0 8 9 】

Example 12: Amplification step for a sample

In cases where the amount of a sample is limiting to the invention, the sample material can be amplified by the following approach. In a first step a plurality of mRNAs is treated as described in Example 11 to replace the cap structure by an appropriate oligonucleotide having a class I Is recognition site. In a second step the aforementioned template is amplified by a PCR step using a primer complementary to the linker and a poly-A primer. The PCR product can be used for the invention as described in the Examples 1.

【 0 0 9 0 】

Example 13: Utilization of extended 5'-end sequences

Initial 5' end sequences obtained for concatamers can be used to synthesis sequencing primers to obtain extended sequence information on the 5' end of a transcribed region.

【 0 0 9 1 】

Example 14: Gene inactivation

Sequence information obtained from 5' end specific sequence tags can be used for the design of anti-sense probes, which could be applied in knockdown studies.

【 0 0 9 2 】

【発明の効果】

By the present invention, novel means by which not only the information on the nucleotide sequences of mRNAs contained in a sample may be obtained

ined, but also novel genes could be cloned. By the method of the present invention, information on the nucleotide sequences of the 5' end regions of a plurality of nucleic acids said mRNAs and cDNAs in the sample could effectively be obtained. Since the information on the nucleotide sequences of the 5' end regions is obtained, unknown genes can be cloned after the identification of a novel transcript. Further, it may be possible to attain mapping of transcription start sites, mapping of promoter usage pattern, analysis of SNPs in promoters, creating gene networking by combining the expression analysis, alternative promoter usage and the other data in this disclosure, and selective recovery of promoter regions in fragmented genomic DNA.

【 0 0 9 3 】

In particular, the invention has a great impact on identification, cloning and further analysis of promoter regions. After sequencing concatamer libraries holding information on a plurality of 5' ends, a statistical analysis on the distribution on the transcriptional start sites will be possible. Changes between different physiological conditions switch the mRNA transcription machinery into new "status". Such a "transcriptional status" can be measured by computing (1) the presence of the transcription starting points, (2) the digital expression of the various transcriptional factors by counting their expression by counting the tags, and correlating the presence of starting point, the transcription factors. More information will be obtained on the gene networking by comparing the perturbation of gene expression between two different conditions. Such comparisons of transcriptional conditions between various disease and normal tissues could allow for the design of new and very comprehensive diagnostic tools. Thus the invention will be of high commercial value in gene discovery and gene analysis, and it is envisioned that the invention will be of use in the development of novel diagnostic and therapeutic products.

cts.

【 0 0 9 4 】

[REFERENCES]

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW, Serial analysis of gene expression, Science 1995 Oct 20;270(5235):484-7

US patent 5,866,330 (SAGE)

US patent 5,695,937 (SAGE)

Piero CARNINCI et al., METHODS IN ENZYMOLOGY, VOL. 303, pp. 19-44, 1999

Lee S, Clark T, Chen J, Zhou G, Scott LR, Rowley JD, Wang SM, Correct identification of genes from serial analysis of gene expression tag sequences, Genomics 2002 Apr;79(4):598-602

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE, Using the transcriptome to annotate the genome, Nat Biotechnol 2002 May;20(5):508-12

Maruyama K and Sugano S, Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene. 1994, Vol. 138:171-4

Edery I, Chu LL, Sonenberg N, Pelletier J, An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture), Mol Cell Biol 1995 Jun;15(6):3363-71

US patent 6,022,715 (GenSet)

Shibata Y, Carninci P, Watahiki A, Shiraki T, Konno H, Muramatsu M, Hayashizaki Y, Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method, Biotechniques 2001 Jun;30(6):1250-4

Sambrook J and Russel DW, Molecular Cloning A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York, 2001

Carninci P, Shibata Y, Hayatsu N, Itoh M, Shiraki T, Hirozane T, Watahiki A, Shibata K, Konno H, Muramatsu M, Hayashizaki Y, Balanced-size and l

ong-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis, Genomics. 2001 Sep;77(1-2):79-90.

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA, Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL, Nucleic Acids Res 1998 Jan 1;26(1):362-7

Maruyama K, Sugano S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene. 1994 Jan 28;138(1-2):171-4.

Jordan B., DNA Microarrays: Gene Expression Applications, Springer-Verlag, Berlin Heidelberg New York, 2001

Schena A, DNA Microarrays, A Practical Approach, Oxford University Press, Oxford 1999

US patent 5,962,272 (Clontech)

Carninci P, Shiraki T, Mizuno Y, Muramatsu M, Hayashizaki Y, Extra-long first-strand cDNA synthesis, Biotechniques 2002 May;32(5):984-5

US patents 6,352,828; 6,306,597; 6,280,935; 6,265,163; 5,695,934 (Lynx)

#### 4 Brief Description of Drawings

##### 【図 1】

An example for the preparation of a plurality of 1<sup>st</sup> strand cDNAs is presented, where the starting material can be RNA derived from a biological sample or a cDNA library.

##### 【図 2】

An example for the cloning of 5' -end specific tags into concatemers is presented. The example is including but not limited to the use of the restriction enzymes Gsu I, Bgl II and Eco RI.

##### 【図 3】

An example for a 1<sup>st</sup> linker to be used for the cloning of 5' -end speci

fic tags is presented. The example is including but not limited to the use of the restriction enzymes Bgl II, Gsu I, and Mme I.

【図 4】

An example for a 2<sup>nd</sup> linker to be used for the cloning of 5' -end specific di-tags is presented. The example is including but not limited to the use of the restriction enzymes Bgl II, Gsu I, Mme I and Eco RI.

【図 5】

An example for the structure of a di-tag is presented. The example is including but not limited to the use of the restriction enzymes Bgl II, Gsu I, Mme I and Eco RI.

【図 6】

An example for the use of a 5' -end specific linker is presented, in which the linker is used for the enrichment of individual nucleic acids and their sequencing.



【書類名】

外国語図面

【図 1】

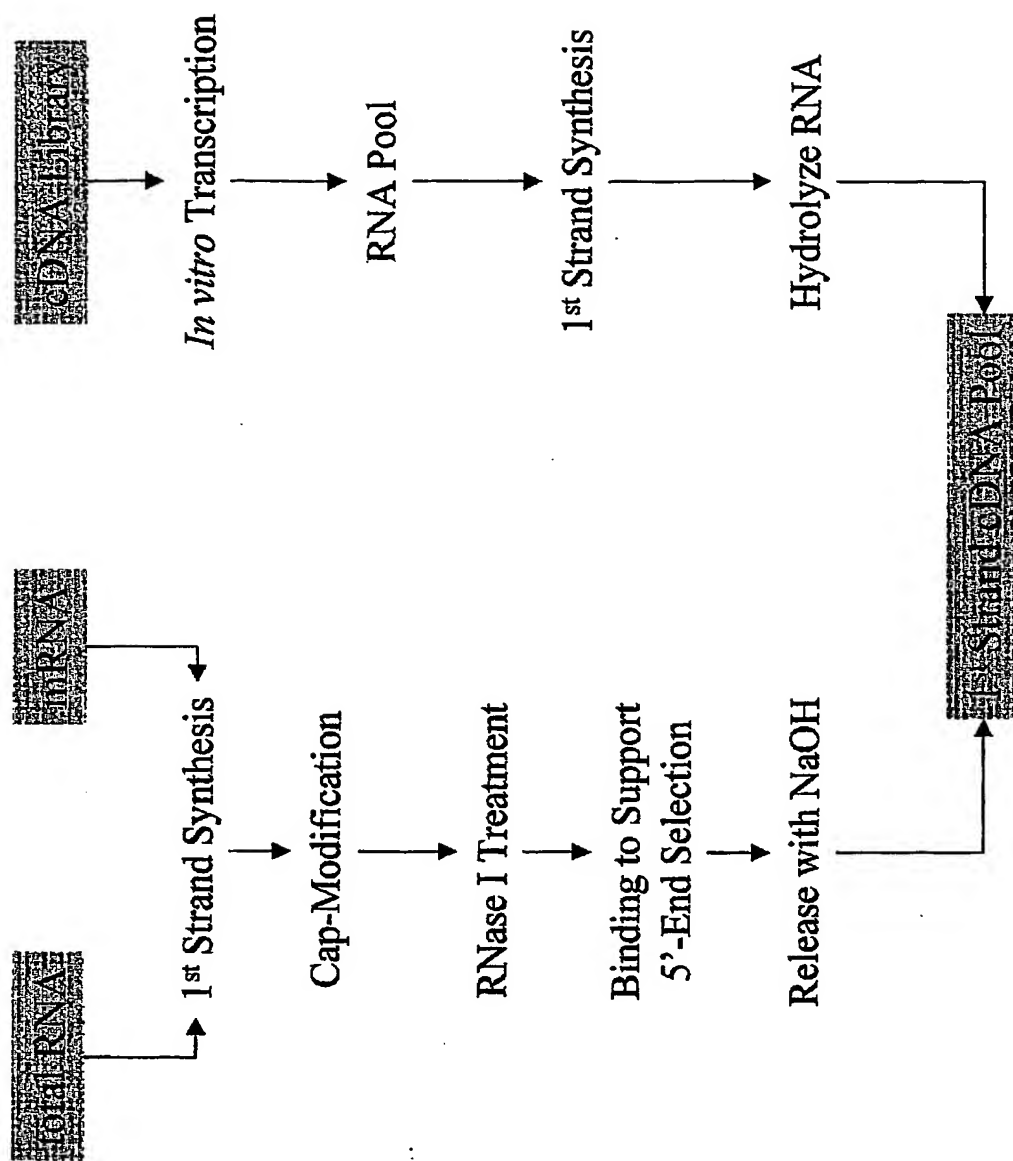


Fig.1: 1<sup>st</sup> Strand Synthesis

【図 2】

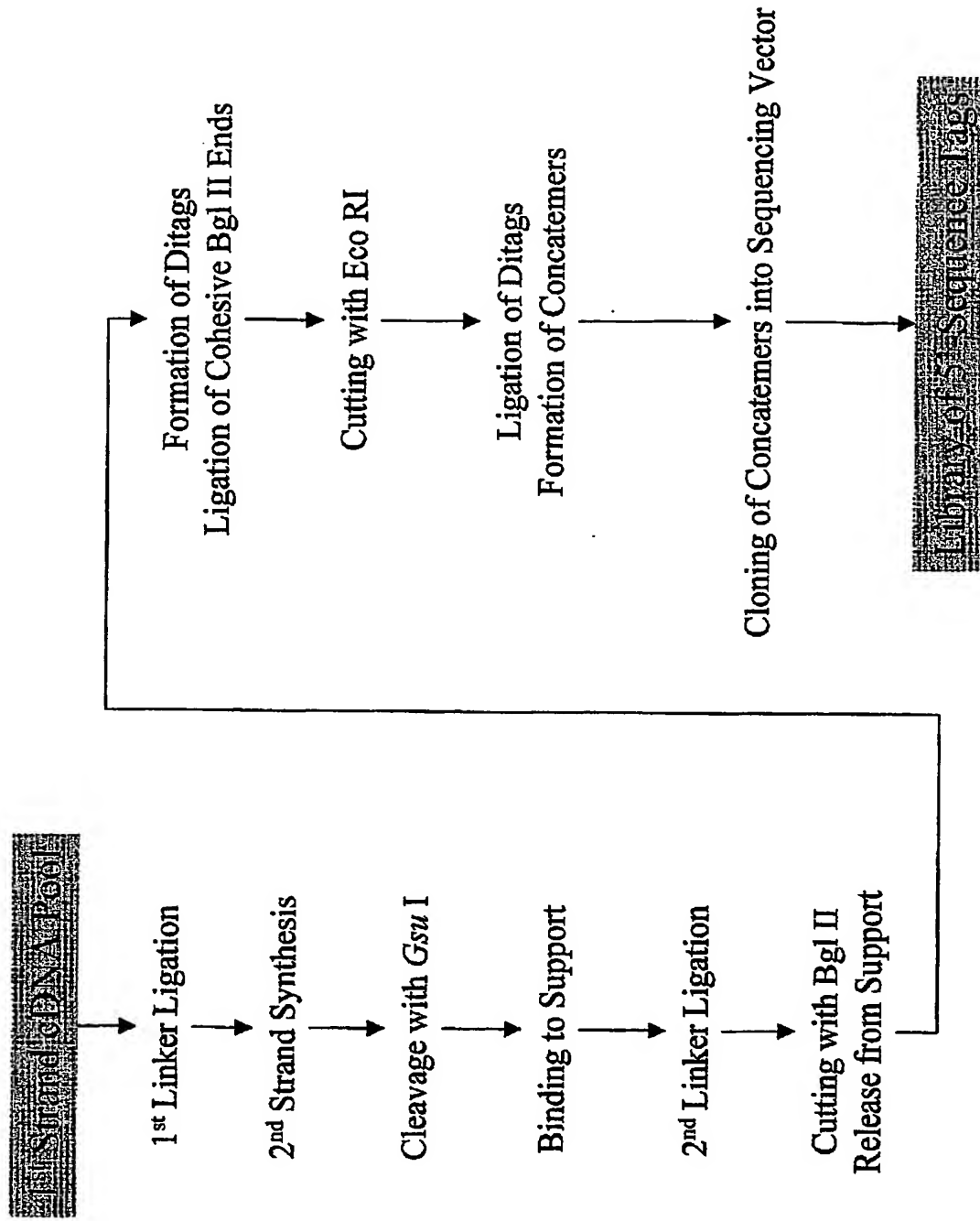
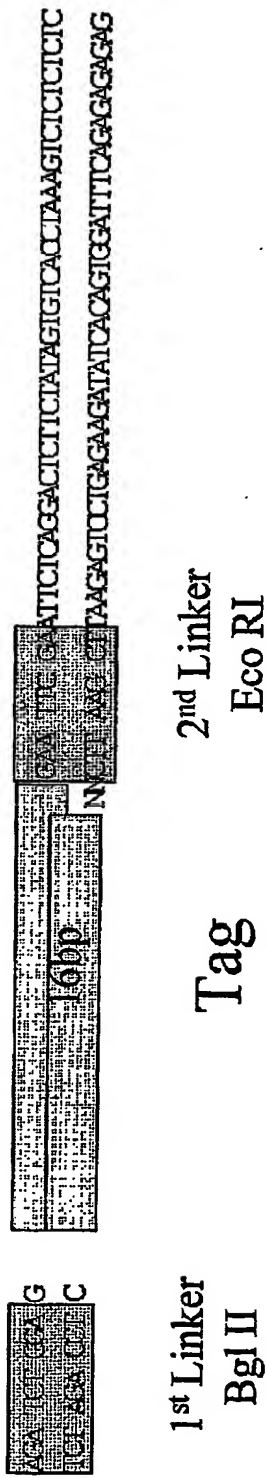


Fig.2: Cloning of Concatemers

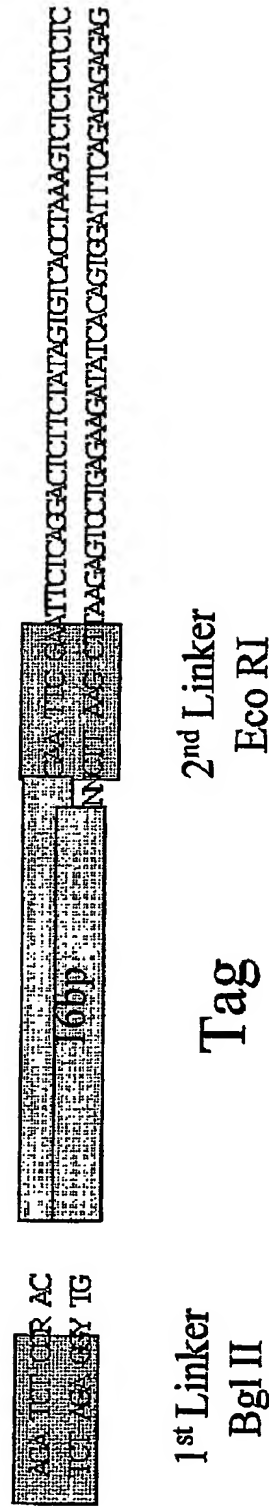


【図 4】

For the use of *Gsu* I:



For the use of *Mme* I:

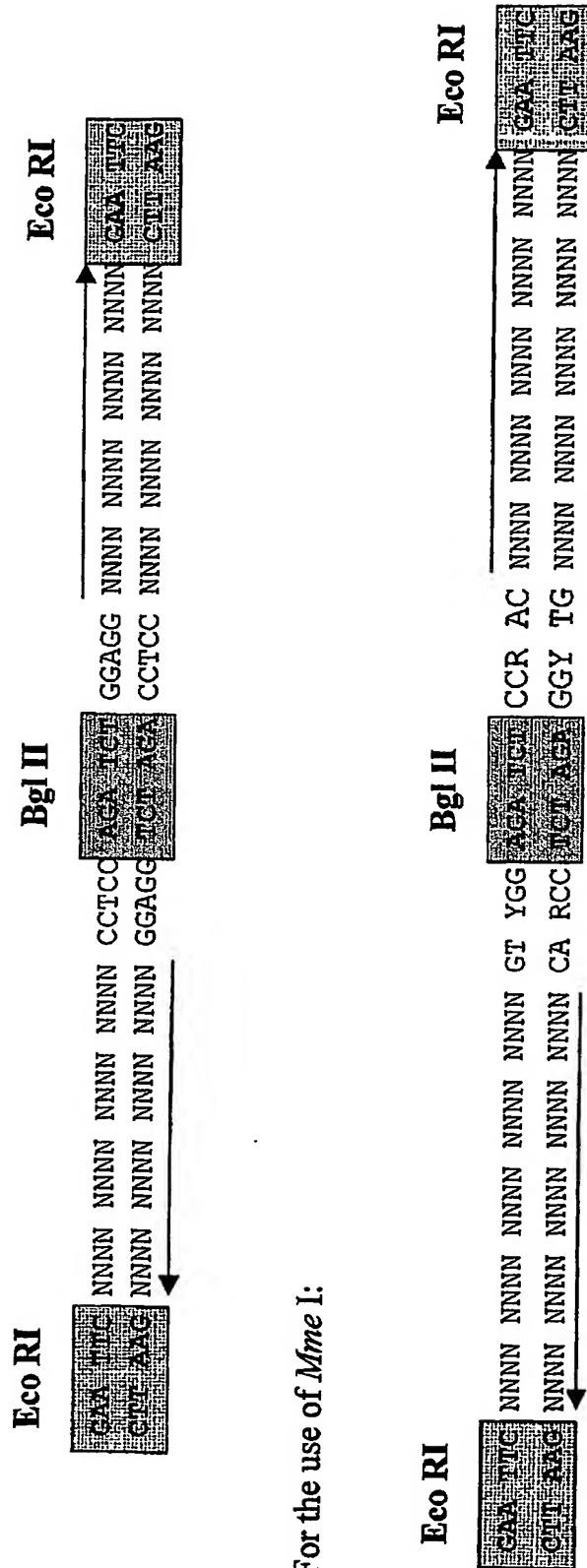


R = G or A  
Y = C or T

Fig.4: 2<sup>nd</sup> Linker Ligation

【図 5】

### For the use of Gsu I:



R = G or A  
Y = C or T

**Fig.5: Structure of a Di-tag**

【図 6】

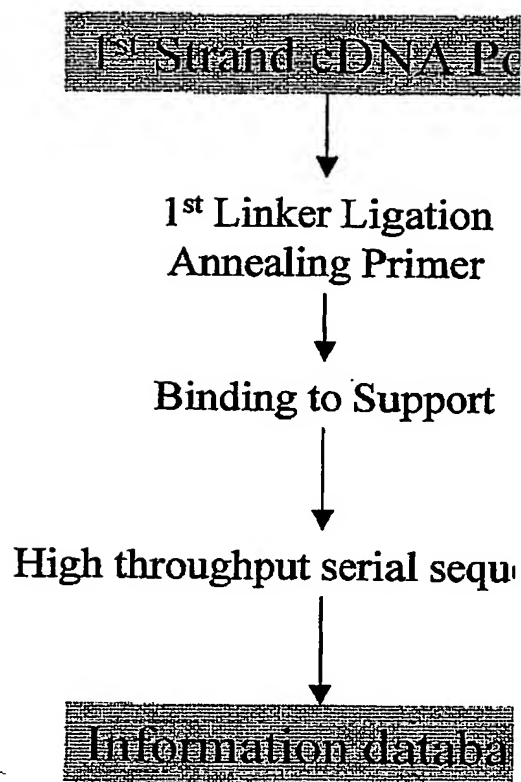


Fig.6: Serial Sequencing

## 【書類名】 外国語要約書

## 1 Abstract

A method is disclosed to obtain the 5' ends of transcribed regions from a plurality of nucleic acid fragments obtained from biological materials or synthetic pools. DNA fragments encoding the 5' ends are enriched for their individual analysis or for the analysis of concatamers thereof. The sequence information derived from 5' ends can be used for characterization and cloning of the transcriptome.

## 2 Representative Drawing None

## 認定・付加情報

特許出願の番号	特願 2002-171851
受付番号	50200855527
書類名	特許願
担当官	森吉 美智枝 7577
作成日	平成14年 9月25日

## &lt;認定情報・付加情報&gt;

【提出日】	平成14年 6月12日
【特許出願人】	
【識別番号】	000006792
【住所又は居所】	埼玉県和光市広沢2番1号
【氏名又は名称】	理化学研究所
【特許出願人】	
【識別番号】	501293666
【住所又は居所】	東京都港区三田1丁目3番35号
【氏名又は名称】	株式会社ダナフォーム
【代理人】	申請人
【識別番号】	100088546
【住所又は居所】	東京都千代田区飯田橋4丁目5番12号 岩田ビ ル6階 谷川国際特許事務所
【氏名又は名称】	谷川 英次郎

次頁無



特願2002-171851

出願人履歴情報

識別番号

[000006792]

1. 変更年月日

1990年 8月28日

[変更理由]

新規登録

住 所

埼玉県和光市広沢2番1号

氏 名

理化学研究所

特願 2002-171851

出 願 人 履 歴 情 報

識別番号

[502049114]

1. 変更年月日      2002年   3月28日  
  [変更理由]      識別番号の二重登録による抹消  
  [統合先識別番号] 501293666  
          住 所      東京都港区三田1丁目3番35号  
          氏 名      株式会社ダナフォーム

特願 2002-171851

出 願 人 履 歴 情 報

識別番号 [501293666]

1. 変更年月日 2002年 3月28日  
[変更理由] 識別番号の二重登録による統合  
[統合元識別番号] 502049114  
住 所 東京都港区三田1丁目3番35号  
氏 名 株式会社ダナフォーム

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☒ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**